

Lunar dreams, again

There are rumblings from the White House about a grandiose vision for human space flight, ahead of President Bush's re-election campaign. If precedent is anything to go by, the results will be discomforting for NASA and for science.

American presidents seeking re-election are inclined to look for 'big ideas' that will re-invigorate their image with the public. Judging by the Washington rumour mill, the current president is no exception. According to several reliable sources, the Bush administration is considering a grand foray into human space flight, perhaps a return to the Moon. It's doubtful that a bold Kennedy-esque declaration is in the works, but history suggests that this is likely to be a bad way to reach for the stars.

The space shuttle and space station were each approved by presidents (Richard Nixon and Ronald Reagan) who faced re-election and the need to keep NASA busy after a big project (Apollo and the shuttle, respectively). Both conditions apply today, combined with an additional pressure: the Columbia accident has forced the Bush White House to make important decisions on the future of human space flight.

Left to his own devices, Bush would no doubt have avoided the subject. He showed little interest in space before becoming president, managing, for example, not to visit the Johnson Space Center in Houston during six years as governor of Texas. If his only motive now is to appear visionary in an election year, or to create employment for astronauts, NASA and the American people should say 'thanks, but no thanks'.

The space station, the last big presidential space initiative, is exhibit A for what can go wrong. Political and financial support for the project has always been lukewarm, which largely explains why Reagan's plan to complete the project "within a decade" has taken twice as long. Meanwhile, the station's rationale has shifted so many times that its purpose is uncertain even as it nears completion.

There is a case to be made for returning to the Moon. Planetary

scientists have ranked a robotic lunar sample return high on their list of desirable missions. If field geologists are someday sent to Mars, the Moon might make a good proving ground. But could they practice on Earth instead? Can robots do the work just as well? Are lunar observatories preferable to orbiting telescopes? Will the US public be inspired by a trip that astronauts took 30 years ago? And would a near-Earth asteroid make a more compelling destination? These matters should be debated thoroughly and honestly before astronauts charge off to the Moon.

Perhaps a more fundamental question is whether NASA should lead that charge. The agency has shown a disturbing resistance to new ideas, and its technical expertise has been eroded in key areas. Sending astronauts to the Moon requires significant, sustained investment — not just to buy hardware, but to attract the kind of engineering talent that flocked to the Apollo programme in the 1960s.

What the space agency needs most right now is a vision that aligns with its budget. The Bush administration's plan may emphasize the contribution of private or other sources of funds, but in the end these funds won't amount to much. So the federal budget proposal announced at the beginning of February will be the best indicator of whether any space-exploration initiative is serious or not.

NASA administrator Sean O'Keefe has been arguing for two years that his agency should develop space technologies without a specific destination such as the Moon or Mars in mind. That never seemed a good idea, just a formula for aimless tinkering. Despite the unfortunate precedents of the shuttle and the space station, the United States does need at least to consider the idea of an overarching goal for human space flight. But a grand mission without the resources to accomplish it would be a recipe for decline — and possibly disaster. ■

A new home for molecular biologists

Celebrating a major relaunch of a Nature journal.

A chronic gap in the spectrum of Nature research journals is now to be filled. Following the launch in 1992 of the first such journal, *Nature Genetics*, we have progressively served several key disciplines in biology with such community-oriented publications. They seem to have fulfilled their initial goals, with all of them achieving a healthy subscription base and most of them lying at the top of their respective categories for impact factor.

But for too long, one group has been left without such a journal: molecular biologists. That label does not do justice to the diversity of interests of those who study biological phenomena at the molecular level. But many of them will, we hope, be pleased to find a new outlet for their outstanding papers in a new title to be launched at the beginning of January: *Nature Structural & Molecular Biology*.

In developing this publication, it has been important to preserve the great strengths of its predecessor, *Nature Structural Biology*, which has had the highest impact factor in its category — the new journal will carry forward its impact factor of 11.8. As its title indicates, the new journal will continue to serve its traditional community

by including in its scope papers whose main message is structural.

But as its launch issue will demonstrate, it will also publish a significant proportion of papers that have no structural component at all, but which fall within the general goal of understanding biological processes at a molecular and mechanistic level. The emphasis of the journal will be on *in vitro* studies. The scope will include processes in the cell nucleus: DNA replication, repair, recombination, transcription, RNA processing and chromatin structure and remodeling. It will also include molecular-level studies of translation, signal transduction and protein folding, processing and degradation (for a complete list, see www.nature.com/nsmb). Analysis beyond a single component in a process and of coupling between cellular processes will also be welcome.

The number of pages in the journal and the number of editors on its staff are both being increased to accommodate this significant extension of its scope. Any author who is interested in submitting a paper to *Nature Structural & Molecular Biology* should visit <http://www.nature.com/nsb/esubmission/index.html>. ■





On the list
Bush administration
under attack over
endangered species
p592



Not guilty
Plague trial triggers
unease among
researchers
p593



Summing up
Demand for
mathematics sparks
institute launch
p594



Sound the horn
Trophy hunters
hamper development
of male sheep
p595

Berkeley accused of biotech bias as ecologist is denied tenure

Rex Dalton, San Francisco

An ecologist at the University of California, Berkeley, best known for his outspoken criticism of genetically modified crops and of the university's links with the biotech industry, has been denied tenure. But some of his colleagues are now questioning the integrity of the decision-making process.

The Berkeley campus has been wracked by dissent ever since it signed a lucrative deal in 1998 with the Swiss-based firm Novartis, giving the company privileged access to the university's plant scientists. Ignacio Chapela was prominent among a group of vocal protesters against the deal. Subsequently, he became embroiled in controversy after publishing disputed research suggesting that transgenes flowed from modified crops into natural maize in his native Mexico.

Chapela's supporters now charge that his denial of tenure calls into question the prestigious university's willingness to back academics who challenge powerful agricultural industrial interests. But university administrators argue that Chapela's publishing record in the seven years since he arrived at Berkeley is too weak to justify tenure.

One Berkeley scientist involved in the tenure review was so upset at the handling of the case that he has broken the strict confidentiality of the process to complain. Population biologist Wayne Getz, who sat on an ad hoc faculty committee that recommended giving Chapela tenure, says that the ecologist received overwhelming faculty support, but alleges that the review then was "hijacked" by Chapela's opponents in the university.

"The process was so irregular; it is illegitimate," asserts Chapela, who received notice on 26 November from Berkeley's chancellor Robert Berdahl that his academic contract will expire next June. University officials won't comment on the specifics of Chapela's case, but a spokesman says: "We stand by



Ignacio Chapela says that he plans to contest a decision not to grant him tenure.

our tenure process; it is the most strenuous in the country."

Documents from the chancellor's office raise a number of concerns about Chapela's performance, including the disputed maize article and his research publications. "The overall assessment of reviewers was that Chapela's good record of teaching and excellent service stood in sharp contrast to a disappointingly modest publication record," says the document rejecting tenure.

Chapela works for the university's Department of Environmental Science, Policy, and Management. Shortly after arriving at Berkeley, he became embroiled in the row over the deal with Novartis — now known as Syngenta. Then, in 2001, Chapela and a student published a paper in *Nature* reporting that transgenes had flowed into native maize in the southern Mexican state of Oaxaca, fragmenting and integrating across the genome (D. Quist and I. H. Chapela *Nature* 414, 541–543; 2001). The article stirred international debate and, after additional review, *Nature* issued a statement saying that it would not have been published had certain technical issues been uncovered during the paper's initial review. Chapela and his student acknowledged some flaws, but stood by their main findings (D. Quist and I. H. Chapela *Nature* 416, 602; 2002).

Chapela's tenure at Berkeley has been under review since November 2000. As part of the process, in his department, 32 faculty members voted for tenure and one against, with three abstentions. And in summer 2002, an ad hoc committee of five colleagues familiar with Chapela's field voted unanimously in favour of tenure.

But the review then took an unusual course. The chair of the ad hoc committee was quizzed by the university hierarchy about his committee's report and its membership; questions were raised about whether two members

were biased. The chair, whose identity has not been revealed, then resigned in the autumn of 2002, disavowing his committee's report. But committee members weren't told that this had occurred.

Getz, a tenured professor in Chapela's department, only learned of what happened to the report of the committee he had served on in June this year. He immediately wrote to the chancellor's office: "I am concerned that the process of tenure evaluation, that works so well in almost all cases, has somehow been tainted or corrupted by those on our campus who belong to the camp that believes Chapela should not be tenured."

Another faculty member involved in the review says: "The process clearly failed when the chairman resigned. This sort of case sends a chill through the community of researchers."

A university spokesman claims that the ad hoc committee's chair resigned after realizing that his committee didn't have the required expertise to analyse Chapela's research, and argues that these actions were compatible with Berkeley's normal tenure-review process.

Chapela and other faculty members were this week planning to hold a meeting on academic freedom at the Berkeley campus. His tenure battle is likely to take centre stage. "I just can't walk away," Chapela says. "I have to fight."

I. CHAPELA

Ecologists attack endangered-species logjam

Betsy Mason, Washington

US ecologists and environmental groups are becoming increasingly frustrated at what they see as attempts by the Bush administration to hobble the Endangered Species Act.

Since the administration took office in 2001, additions to the endangered species list have plummeted to an average rate of just eight per year, compared with 65 species per year during the Clinton years and 59 per year under George Bush, the current president's father, from 1989 to 1992.

Ecologists also say that actual and pending changes in the law could further curtail the effectiveness of the act, which was passed in 1973 to protect threatened animal and plant species.

"A lot of things that this administration has done have been intended to do harm to the Endangered Species Act," says David Wilcove, an ecologist who studies endangered species at Princeton University, New Jersey.

On 3 December, Defenders of Wildlife, a Washington-based environmental group, issued a report arguing that the administration had failed to request enough funding to support the act and refused to comply with court orders, leaving endangered species such as the manatee and the desert tortoise without protection. "We see a very disturbing picture of an administration attacking the Endangered Species Act," says Rodger Schlickeisen, the group's president.

Craig Manson, the assistant secretary at the US Department of the Interior responsible for implementing the act, says that the report doesn't accurately reflect the administration's treatment of endangered species. He says that efforts to protect species by cooperating with private landowners are paying off. "We have put hundreds of millions

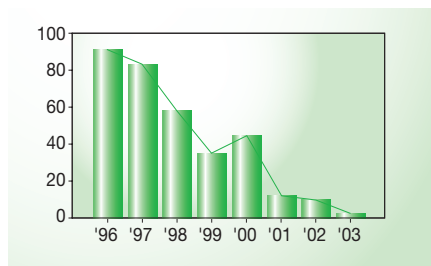


In deep water: government inaction is threatening endangered species such as the manatee.

of dollars into the preservation of habitat with the goal of avoiding listings."

And Manson says that the drop in new species listings (see graph) is caused largely by a rash of court orders from environmental groups requesting protection orders, which he claims have left the Fish and Wildlife Service short of resources for listing new species.

Michael Bean, who chairs the wildlife programme at the Environmental Defense Fund, a New York-based environmental



New species listed each year under the Endangered Species Act.

group, agrees that animals such as the northern aplomado falcon in Texas and the nene in Hawaii have benefited from government cooperation with private landowners. "Progress is being made on the ground," he says.

But the news is not all good, says Bean. "There is a concerted effort on the part of the Department of the Interior to diminish provisions of the act," he says.

And, ecologists say, new legislation is set to accelerate the trend. Under the Healthy Forest Restoration Act, for example, which was signed by President Bush on 3 December, federal agencies such as the US Department of Agriculture's Forest Service are no longer required to consult with biologists from the Fish and Wildlife Service before carrying out actions that could affect endangered species.

"This could lead to a situation where harm could be done to endangered species under the radar of the Fish and Wildlife Service," says Wilcove.

An amendment to the Endangered Species Act, introduced in the House of Representatives in April and supported by the Bush administration, would require that only empirical data and peer-reviewed research be considered for species listings and critical habitat designations, rather than the current use of the "best available science".

This would make rare or elusive animals that have received little scientific attention ineligible, says ecologist Gordon Orians, former president of the Ecological Society of America. "The idea that you could determine where a population is headed based solely on empirical data without a population model is ridiculous," he says.

"There are a lot of species in big trouble in this country, and they are not all listed yet," says Orians. "It's so discouraging what this administration is trying to do."

Swamp row bogs down Singapore's bid to reclaim land

David Cyranoski, Singapore

Marine biologists are being asked to assess the impact of a land-reclamation project off Singapore which, neighbouring Malaysia charges, could damage mangrove swamps and other important ecological features.

In September, Malaysia appealed to the Hamburg-based International Tribunal for the Law of the Sea (ITLOS), asserting that the project infringes on its territory and that Singapore has failed to undertake the environmental assessment required by the United Nations Convention on the Law of the Sea.

And last month, the tribunal ordered the

two countries to assemble a group of independent experts to assess the impact of the 33-square-kilometre project, which began three years ago and aims to expand the island of Pulau Tekong, which sits near the mouth of Malaysia's Johor river.

The group is due to report by next October. But in the meantime, some experts are already criticizing the project. Ving Ching Chong, an ecologist at the University of Malaya in Kuala Lumpur, says that it will probably damage mangrove forest at the mouth of the river. He says that the swamps are valuable spawning grounds for marine life, and protect the

land behind them from storms.

Chong adds that the project, by constricting the area for water flow, will increase the tidal flow and river current, and wash away the soil in which the trees stand. "They won't be able to take it. They'll just topple over," he says.

But a Singaporean marine biologist, who declined to be named, says that the project is well under way and has had no impact on the mangroves.

Other Singaporean researchers allege that Malaysia's real gripe about the project concerns the access of ships from its ports to shipping routes in the area's crowded seas.

Plague trial verdict leaves biologists split on biodefence

Erika Check, Washington

On 1 December, a jury in Lubbock, Texas, split the difference in its verdict on the celebrated case of plague researcher Thomas Butler. The jury cleared Butler of serious crimes relating to his handling of samples of plague bacteria, but convicted him of defrauding his university through contracts with drug companies.

In the aftermath of the case, the research community is likewise split over what its ultimate impact will be. Some insist that what they see as a vindictive prosecution of an esteemed researcher will scare many biologists away from sensitive but important work. Others say it that the case has not harmed biology — and should not alarm researchers who take care to follow the rules.

"Butler [was] innocent of the initial FBI charges relating to handling the select agent but found guilty because of peripheral information they dug up after the fact," says Stanley Falkow, a microbiologist at the University of Stanford, California, who has spoken out against the prosecution. "I'm not sure that this will calm anyone's fears — certainly not mine."

But others point out that such worries have not prevented researchers from scrambling to apply for the funds that the federal government has offered for biodefence work. Ronald Atlas of the University of Louisville in Kentucky, a former president of the American Society for Microbiology, says that many scientists are eager to work on such projects.

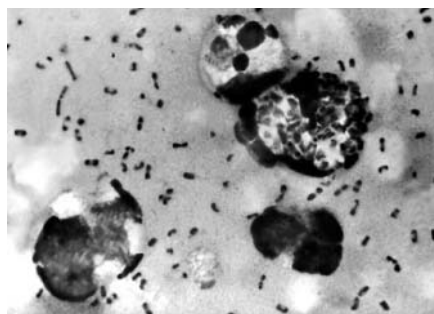
Atlas acknowledges that scientists must comply with a new array of regulations in the wake of the postal anthrax attacks in 2001. "We live in a new regulatory environment; it is our responsibility as scientists and citizens to comply with the laws and regulations," Atlas said in an e-mail. "I have not seen evidence that [the regulations] are having an adverse impact." He declined to comment specifically on the Butler case.

Butler was arrested in January after he reported that vials of plague bacteria were missing from his laboratory at Texas Tech University in Lubbock. He was charged with 15 crimes, including lying to federal agents about the vials, transporting them without the proper paperwork, and cheating on his taxes. Later, the Department of Justice charged him with 54 further crimes relating to his "shadow contracts" with drug firms.

The jury acquitted Butler of almost all of the 15 original charges — except shipping the bacteria overseas illegally — but convicted him of the fraud charges relating to his drug-company contracts. Federal judge Sam



Researchers are unsure whether to draw comfort from the acquittal of Thomas Butler on most charges of illegally transporting plague (below).



Cummings has until mid-February to set Butler's sentence. Butler's lawyers say that they will appeal against the convictions.

Regardless of the appeal outcome, Butler faces an uncertain future. On 12 December, officials at Texas Tech will meet to resume an administrative hearing that could result in his dismissal. The hearing began before Butler's trial and stemmed from disagreements between Butler and the university over financial and administrative issues. And already this year, Butler and his family have spent almost \$1 million on legal fees, his lawyer says.

But Peter Agre, a joint winner of this year's Nobel Prize in Chemistry, says that Butler should be able to continue his research career elsewhere because most scientists are sympathetic to him. "I believe that Dr Butler will survive with his name intact, but at a high cost," Agre says. "It's been a very emotionally draining year."

National Science Foundation facing budget let-down

Geoff Brumfiel, Washington

Science advocates in the United States are disappointed at a new spending bill that will give the National Science Foundation (NSF) far less cash than was recommended in a law passed just a year ago.

The omnibus appropriations bill, which sets 2004 funding levels for much of the government, provides a 5% increase to the NSF's total budget of about \$5.3 billion. That is far less than was recommended in a bill signed by President George Bush last December, which suggested doubling the agency's budget by 2007, by awarding 15% increases each year (see *Nature* 420, 257; 2002).

"Those of us who had worked on the authorizing bill would have liked to see at least a double-digit increase," says Samuel Rankin, chair of the Coalition for National Science Funding, one of the original advocates of the 2002 legislation.

Congressional staff who worked on the 2002 bill are also disgruntled at the smaller increase set out in the 2004 budget, but say that the plan to double the NSF budget was only a suggestion. "It's not like when the [2002] bill passed we thought it would be a linear progression," says David Goldston, chief of staff at the House Committee on Science, which helped to draft the legislation. He points out that many specific initiatives supported by the committee were funded in the 2004 budget bill.

The NSF was not the only research agency to receive lacklustre funding in the 2004 budget, which was passed by the House on 8 December and is expected to be passed by the Senate early next year. The National Institutes of Health will receive an increase of just 3.2%, to about \$27 billion. Janet Shoemaker, head of public affairs at the American Society for Microbiology in Washington DC, says she doubts that the president's budget proposal for 2005, to be released in February, will be much more generous.

In the 2004 bill, NASA's budget was held at \$15 billion, but money for research was trimmed by 0.4%, partly because of budget pressure arising from the loss of the shuttle Columbia in February. The research budget of the National Institute of Standards and Technology also fell by 4%, to \$506 million. The Department of Energy's Office of Science fared better, winning a 3.8% increase to \$3.2 billion. Much of that money will go towards research on nuclear fusion, computing and biology.

Mystery remains as journal withdraws paper

John Whitfield, London

A mathematics journal has withdrawn a paper that claimed to crack one of the discipline's great mysteries after reviewing and accepting the work and publishing it online.

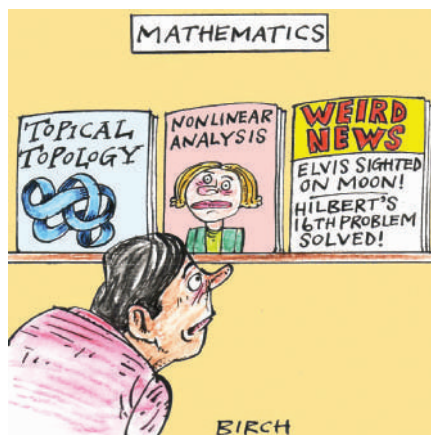
On 18 November, *Nonlinear Analysis* published a paper by Elin Oxenhielm — a postgraduate student in mathematics at the University of Stockholm, Sweden — which presented itself as a solution to the second part of Hilbert's sixteenth problem, one of a set of challenges laid out by German mathematician David Hilbert in 1900.

If a solution were validated, mathematicians agree, it would be a significant step towards a complete solution to the problem. Oxenhielm predicts just that: "We could find one in a year or so, if we're lucky," she says.

The work was described in a 24 November press release from Oxenhielm and covered in several media outlets including the BBC. But the paper immediately came under fire from mathematicians. "It's completely inadequate — I can't imagine who would have thought it was a proof," says John Mather of Princeton University, New Jersey.

Critics include Oxenhielm's supervisor, Yishao Zhou, who put a statement on her website saying: "The paper is incomplete and contains serious mistakes."

Hilbert specified 23 problems that he said should drive mathematical research. Solving any one of them is almost guaranteed to



make a mathematician's name, and by 2000 all but three had been solved.

The sixteenth, the problem of the topology of algebraic curves and surfaces, deals with the territory where geometry meets algebra. Its second part involves showing that the number of periodic solutions to a differential equation is finite.

Such periodic solutions are also known as limit cycles — stable, oscillating trajectories to which a system will return if perturbed. Limit cycles are common in nature, and a proof of the second part could lead to a better understanding of heartbeats, animal movements and the kind of runaway vibrations that can shake a structure to bits.

Oxenhielm formulated her proof using

'describing functions' — which can predict roughly the presence of limit cycles in non-linear equations.

A few minutes' scrutiny is enough to show that her reasoning is false, says mathematician Grigori Rozenblium of Chalmers University of Technology in Gothenburg, Sweden. The approximate solutions studied by Oxenhielm cannot provide the exact answers demanded in a proof, he says, and some of her equations contain exact terms where approximate ones should be used.

The work should never have been published, Rozenblium says: "It's impossible to understand the behaviour of the journal, which is one of the leaders in its field."

Nonlinear Analysis pulled the paper on 4 December. "Publication has been halted until a thorough investigation into the matter has been handled," says editor-in-chief V. Lakshmikantham, a mathematician at the Florida Institute of Technology in Melbourne.

Originally approved by one reviewer, the paper has now been sent to two more mathematicians for further round of review, along with a defence by Oxenhielm, who says that the critics do not understand her methods.

She refuses to comment further. "*Nonlinear Analysis*' editors have evaluated the paper, they accepted it for publication and they have the copyright of its contents — and thus they are responsible for its correctness," she told Norwegian newspaper *Aftenposten*. ■

Maths institute planned to meet multiplying demand

Jim Giles, London

Ambitious plans are to be unveiled this week for a top-level mathematics institute at Imperial College London.

The Institute of Mathematical Sciences, due to open in early 2005, will host six applied research groups, each of which will explore a single theme for an initial period of three to five years. Every group will be headed by an existing Imperial researcher, who will employ five or six postdoctoral researchers and visiting scientists.

Phil Hall, a mathematician at Imperial who will become the institute's first director, says it is being created in response to growing demand for mathematicians from other departments in the university. He cites the study of protein folding as an example of an area that is increasingly involving maths, and says that likely areas of interest for the institute include finance, biostatistics and string theory.

"We want to increase the quantity of this type of work at Imperial by an order of magnitude," says Hall.



Count on it: £3 million will help spruce up Imperial's new mathematics centre, above.

Imperial's plan, which was due to be announced on 11 December, is conceptually similar to established US centres backed by the National Science Foundation, such as the

Mathematical Sciences Research Institute in Berkeley, California. Established in 1982, the Berkeley institute employs more than 20 postdoctoral fellows. "There is an emphasis on bringing young people together with experienced researchers" at the US centres, says Chris Sogge, chair of the mathematics department at Johns Hopkins University in Baltimore, Maryland.

British mathematicians have welcomed the plan, although some have questioned whether the university will be able to attract enough grants from research agencies to sustain the 50 staff that it plans to have.

But Hall says that by working in interdisciplinary areas, the institute will be able to tap into funding sources to which mathematicians on their own rarely have access, such as the Wellcome Trust, Britain's largest medical charity. Nonetheless, he adds, Imperial expects to subsidize its first three years of operation, in a building to be refurbished with a £3-million (US\$5-million) grant from the Higher Education Funding Council for England. ■



Thinking big: trophy hunters are removing rams with large horns from bighorn sheep populations.

Sheep horns downsized by hunters' taste for trophies

John Whitfield, London

The horns of some bighorn sheep are getting smaller, because hunters are picking off the most impressive rams before they reach their breeding peak.

A study of one sheep population in Canada shows that hunting can harm the gene pool of a species over just a few years. That means there should be tougher restrictions on what animals can be taken, says David Coltman of the University of Sheffield, UK. "For selection to be having this effect is of fundamental importance," he says.

Biologists have long suspected that hunting can affect animal evolution. Elephant poaching, for example, is thought to have led to an increase in the number of tuskless animals in Africa. And in Canada, the hunting of moose seems to have resulted in animals with smaller antlers.

To pin the relationship down, Coltman and his colleagues studied the sheep of Ram Mountain, Alberta. This Canadian province is home to the world's biggest bighorn sheep, and is a magnet to hunters. Since 1975, 57 of Ram Mountain's rams have been shot — about 10% of all males in the population each year from 1975 to 1996. In 1996, the government restricted hunting to males with a large, 'full curl' of horns, which has reduced the cull to zero in recent years. Coltman looked at rams from 1971 to 2002, and found that horn size fell by about a quarter over this period (see page 655). Despite the recent drop-off in hunting, horn size has not recovered.

Large horns are generally correlated with

large, healthy rams, says Coltman, so the effect on the population's genetics is probably deeper than the effect on horns alone. He suspects that hunting is also influencing mating behaviour, with fewer rams butting heads to fight for partners.

One reason for the change is that hunters prefer rams with large horns, as they make for more impressive trophies. But it could also be an accidental side effect of some hunting regulations. Restricting hunting to males with large horns is meant to limit the killing of animals that are not old enough to breed, but it also encourages the culling of animals that grow large horns early in life. "You force every hunter to harvest the very animal that you're trying to grow," says Kevin Hurley, a wildlife biologist with the Wyoming Fish and Game Department and executive director of the Northern Wild Sheep and Goat Council.

A better strategy may be to limit the number of hunting permits, argue both Hurley and Coltman. In the western United States, it is common to offer a limited number of permits for bighorn sheep by lottery or auction — the right to shoot a single ram frequently fetches as much as \$100,000.

Hunters are generally sympathetic to the need for management, says Kelly Semple, executive director of Hunting for Tomorrow, a coalition of hunting groups based in Edmonton, Alberta, as hunters do not want to drive large-horned animals into extinction. But she warns against generalizing Coltman's results to all species and locations. ■

Europe dithers over regulations for stem-cell research

Alison Abbott, Munich

Europe's debate on research using embryonic stem cells has ended for now in stalemate, leaving cell biologists uncertain as to where they stand.

The research ministers of the European Union's member states met on 3 December but failed to agree on whether European Union funding should be available for research using newly derived human embryonic stem cells. A moratorium on such funding will expire on 31 December.

In the absence of an agreement, the European Commission is now, at least in theory, free to fund future projects in its Sixth Framework Programme that exploit new cell lines. Although some of the European Union's member states are opposed to the use of new cell lines, the commission has said that it will think carefully about individual project proposals on a case-by-case basis.

In January, European Commission officials said that participants in the Sixth Framework Programme could use only embryonic stem cells that had been derived before the end of 2002, and gave the European Parliament and the Council of Ministers a year to develop clear rules on whether new cell lines produced from spare embryos from *in vitro* fertilization clinics could also be used.

Some countries, including the United Kingdom, allow such research to be carried out with public funds, whereas other nations, such as Germany, allow only existing cell lines to be used.

Last month, the European Parliament voted in favour of the use of newly derived cell lines, which some commissioners hoped would encourage waverers within the council also to vote liberally. But in the absence of a decision by the ministers, the European Commission will have to make its own rules about what it will allow in its next call for Framework proposals in June.

Ireland, which will take over the rotating presidency of the European Union from Italy in January, may call for another vote in the spring, but is under no obligation to do so. "Everything is still as open as it ever was," says Octavi Quintana Trias, head of life sciences at the commission's research directorate. ■



Fiery protests: student demonstrations have been raging on Germany's streets this winter.

Standing room only as German protest lectures go on and on

Munich Passengers on public transport in Berlin last week found themselves with something besides adverts to keep their minds occupied. To protest against budget cuts and poor conditions at the city's three main universities, thousands of striking students and professors held courses in crowded subway compartments, as well as on the capital's icy streets.

Students and faculty of the Humboldt University's physics department displayed particular creativity and stamina. They organized a 72-hour marathon physics lecture on Potsdamer Platz — possibly the longest in the history of the field.

Universities throughout Germany are suffering from a crisis in public finances (see *Nature* 426, 372; 2003), but the situation in the capital is particularly grave. Many universities are calling for the introduction of tuition fees, although Germany's Social Democrat–Green government and many students strongly oppose the idea.

California gives red light to sale of transgenic GloFish

Washington The light of a transgenic fish designed as a novelty fluorescent pet will not shine in California in the coming year.

GloFish — a zebrafish that has sea anemone genes stitched into its genome — was created by Yorktown Technologies, a company based in Austin, Texas (see *Nature* 426, 372; 2003). The company applied for an exemption to a California law that restricts the sale of transgenic aquatic animals. But on 3 December, the state's Fish and Game Commission rejected the exemption by three votes to one.

The commissioners largely agreed with scientists who say that the GloFish will probably not damage the environment if dumped into the wild. But they also concluded that aesthetic reasons are not sufficient justification for the genetic modification of animals.

GloFish's detractors have welcomed the decision but remain concerned that the US Food and Drug Administration (FDA) has refused to consider banning the GloFish outright. They plan to press their case with the FDA in a meeting on 15 December.

Yorktown Technologies plans to start selling the GloFish in every state except California on 5 January.

Cancer centre finally finds its head

Munich Otmar Wiestler, a neuropathologist at the University of Bonn, was last week appointed head of one of Germany's premier biomedical research institutes, the German Cancer Research Centre in Heidelberg.

The appointment by the German cabinet ends a nine-month search for a successor to Harald zur Hausen, who retired in March. A first attempt in April failed after the designated new director, virologist Bernhard Fleckenstein, unexpectedly stood down at the last minute (see *Nature* 423, 214; 2003).

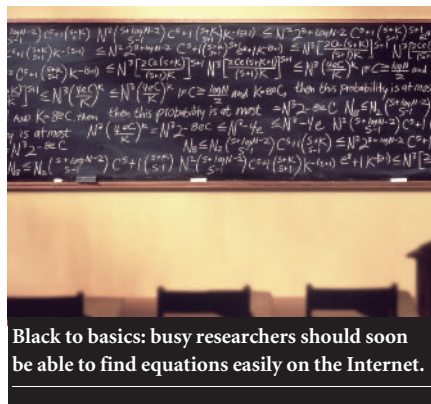
Wiestler sparked intense debate in Germany about the ethics of stem-cell research three years ago, when he applied for public funds for research involving human embryonic stem cells.

He says that his main task will be to bring discoveries closer to clinical application. "Biomedical research will remain at the core of our activities," he says, "but the centre will also need to expand its collaboration with universities, hospitals and industry."

Mathematicians get fresh formula for web searches

Washington Searching for mathematical formulae on the Internet could soon be as simple as looking for words.

Most web pages are coded in hypertext markup language (HTML), which contains information about a page's appearance, including the text that appears on the screen. Search engines tend to trawl through this information when looking for keywords or phrases. But they usually miss equations, as these are often pasted onto web pages as images that cannot be searched. This



Black to basics: busy researchers should soon be able to find equations easily on the Internet.

problem can be solved by adding an extra layer of information to a web page, using MathML, a language that describes the symbols used in equations and spells out the meaning of equations as a whole.

Researchers will gather at the University of Minnesota next April to discuss how best to make a search engine based on MathML. Such a tool would help researchers discover when equations from their own field are used to describe other phenomena. The project is funded by the US National Science Foundation and led by Design Science, a company based in Long Beach, California.

► www.ima.umn.edu/complex/spring/Searching.html

Nobel protester gets Swedish physics gong

New York A two-month protest over this year's Nobel Prize in Physiology or Medicine culminated with two ceremonies on opposite sides of the Atlantic this week.

Physicists Paul Lauterbur and Peter Mansfield, who won the Nobel Prize in October for their work on magnetic resonance imaging (MRI), were due to receive their prize in Sweden on Wednesday.

Meanwhile, New York-based physician Raymond Damadian, who claims to have invented MRI and has been campaigning for a share in the prize (see *Nature* 425, 648; 2003), has been given an alternative award. A Swedish inventors' group called Idéforum has offered him a gold medal in physics and technology, to be presented on the same day as the Nobel ceremony. "It's not the Nobel but it's sweet of them to come over here," Damadian says.

The Nobel committee says the decision about the prize will not be changed, because it was carefully researched and because strict statutes forbid it. Damadian says he has not decided whether to continue his protest.

Korea strategy aims to give good scientists a job for life

Tokyo South Korea is hoping to make science a lucrative and long-lasting career choice.

The Ministry of Planning and Budget last week announced a 50% increase in funding for its science and technology universities to 224 billion won (US\$190 million) for next year, featuring more money for scholarships.

"We are having trouble recruiting students into science and technology courses," says Taeho Bark, vice-chancellor for international affairs at Seoul National University. The extra funds should help to solve that, he says.

South Korea also hopes to hang on to experienced researchers for longer. The ministry says it will offer two-year extension contracts to exemplary scientists to encourage them to remain in their posts beyond the mandatory retirement age of 65 years.

Boom, or bust?

Hundreds of millions of dollars are pouring into US biodefence research. You might expect scientists working on infectious diseases to be unequivocally delighted. But things aren't that simple, says Erika Check.

Three years ago, Karl Hostetler founded a company called Chimerix, based on an eminently practical, if unglamorous, idea. Hostetler, who works at the University of California, San Diego, had invented a way to repackage the antiviral drug cidofovir so that it could be taken as a pill rather than intravenously. Faced with 'sexier' investment opportunities, however, venture capitalists weren't much interested, and the company's prospects looked bleak.

But in March 2002, Chimerix's fortunes changed dramatically. At a meeting in Prague, Hostetler and his colleagues revealed

that cidofovir blocks smallpox infections in human cells and lab mice. The US government was searching desperately for smallpox treatments that could counter a potential bioterrorist attack, and giving patients cidofovir pills would be much easier than putting thousands of people on intravenous drips. So in September this year, the National Institute of Allergy and Infectious Diseases (NIAID), based in Bethesda, Maryland, gave Chimerix a five-year, \$36-million-dollar grant to make and test cidofovir pills. "I doubt very much that we would have been able to obtain venture-capital funding to pursue this smallpox

application," says Chimerix vice-president Kevin Anderson. "Given the investment climate of the past couple of years, it would have been virtually impossible."

Chimerix's windfall illustrates one side of the current boom in US biodefence spending. Since the still-unsolved mailed anthrax attacks of October 2001, NIAID has handed out some \$1.8 billion for biodefence research. The money has allowed scientists such as Hostetler and his Chimerix colleagues to pursue work that should make the world a safer place.

But others see signs of trouble in the biodefence bonanza. Some researchers fear that it will distort priorities in infectious-disease research, sucking money away from work to understand and counter natural disease outbreaks that ultimately pose a greater threat to public health. Experts in weapons proliferation, meanwhile, are concerned that the expansion of labs working on potential bioweapons agents will increase the risk of these pathogens getting into terrorists' hands. And many microbiologists are confused and worried by the regulatory framework put in place to reduce this risk — they now fear being dragged through the courts by overzealous federal investigators over an innocent administrative slip-up. Indeed, these worries have been stoked to fever pitch by the prosecution, for alleged breaches of biosafety regulations, of plague researcher Thomas Butler of Texas Tech University in Lubbock (see 'A cause célèbre', left).

A cause célèbre

The sight of a highly respected 62-year-old professor being hauled into a courtroom day after day would be unsettling enough for researchers under any circumstances. But the trial of plague researcher Thomas Butler has profoundly disturbed the US scientific community, which sees him as the victim of a misguided and vindictive government prosecution. Although Butler was cleared on 1 December of the most serious charges laid against him, prominent microbiologists argue that the case will ultimately undermine the US biodefence effort by scaring researchers away from the field.

Butler, who has spent his life studying the bacteria that causes plague, told federal officials in January that some plague samples had gone missing from his lab at Texas Tech University in Lubbock. Investigators who quizzed Butler concluded that he had destroyed the samples and then lied about it, and charged him with making false statements to the FBI.

Butler was cleared of lying to federal investigators, but was found guilty of crimes including theft, defrauding his university and

illegally transporting plague samples overseas. Texas Tech has moved to dismiss him, and he has spent his life's savings on his defence. As *Nature* went to press, he had yet to be sentenced.

Many researchers argue that the rules regarding transport and possession of 'Category A' pathogens — such as the plague bacterium — are so complicated that even experienced scientists such as Butler are confused. They are appalled that Butler was jailed for six days without bail and has endured months of house arrest. In a letter to the Justice Department sent on 15 August, the presidents of the US National Academy of Sciences and the Institute of Medicine wrote that Butler's trial could undermine biodefence research. Four Nobel prizewinning scientists repeated that warning on 3 November, the day on which Butler's trial began. "This respected colleague has been subjected to unfair and disproportionate treatment and the case is having a negative impact on the future of research in this crucial national-security-related field," the laureates said.

Butler's case did not deter researchers from applying for the grants made available for biodefence this year. Much of this money went to researchers who are new to the field. But those who have worked with plague bacteria before have been more cautious. Stanley Falkow, a microbiologist at Stanford University in California, says that he has destroyed his plague samples. On 18 September he wrote to US Attorney General John Ashcroft, arguing that the government's excessive zeal for biosecurity has trampled over Butler's rights, and asking: "How could I possibly permit my students and myself to be subject to the same nightmare if we also made an inadvertent mistake?"



On trial: Thomas Butler outside the courthouse.

Building blocks

So far, the main way in which the influx of biodefence funding has changed the US scientific landscape is by supporting the creation of a network of new labs and centres. A year ago, there were just three maximum-security biosafety-level-4 (BSL-4) labs in the United States, where scientists could work on deadly and infectious airborne pathogens such as the smallpox virus. But thanks to NIAID funding, this capacity will be more than doubled. The agency is building two new BSL-4 labs — one in Fort Detrick, Maryland, on the campus of the US Army Medical Research Institute of Infectious Diseases, which already has its own BSL-4 lab, and the other at NIAID's Rocky Mountain Laboratories in Hamilton, Montana. In September, NIAID gave \$120 million each to Boston University and the



Beating bioterror: the US National Institute of Allergy and Infectious Diseases is spending billions on measures to counteract biological attacks, from training for those first on the scene (left), to research on the agents that cause (below, clockwise from top left) anthrax, plague, Ebola fever and smallpox.

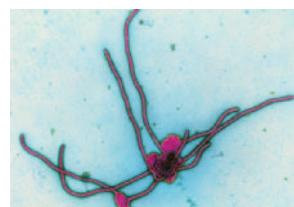
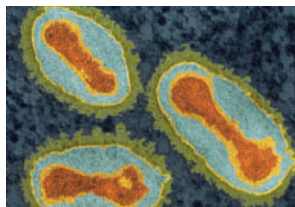
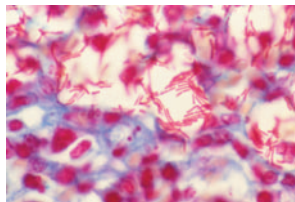
M. KLEINFELD/UPI

University of Texas Medical Branch at Galveston to build two more BSL-4 labs.

These facilities are just part of NIAID's effort to improve infrastructure for bio-defence research. Also in September, the agency gave out eight grants totalling \$350 million over five years to establish a network of regional centres of excellence in biodefence and emerging infectious diseases. These centres will pull together large teams of researchers based at various institutions, supporting everything from basic immunological research to training for 'first responders'—medical personnel who would aid the victims of a bioterrorist attack. NIAID also handed out nine grants of up to \$21 million each to create a network of secure BSL-3 and BSL-2 labs, which will be used to study pathogens such as the yellow fever and West Nile viruses.

NIAID director Anthony Fauci says that the new labs and centres are a crucial element of the federal government's plan to improve the nation's ability to respond to a bioterrorist attack. "The people we've brought together have a commitment that goes from fundamental basic research in pathogenesis, all the way up to and including identifiable and usable countermeasures," he says. "The speed with which we're getting the research off the ground has been quite gratifying."

Meanwhile, NIAID has given individual researchers millions of dollars for projects aimed at helping to devise ways to counter specific disease agents that could be used in a bioterrorist attack. This year, according to estimates compiled by observers of the



agency, NIAID has spent around \$19 million on anthrax alone, and \$12 million on research into poxviruses. Together, these pathogens absorbed half of the total bio-defence funding given out last year, excluding spending on new labs and centres. Federal officials have classified these pathogens, together with the group of viruses that cause Ebola and other haemorrhagic fevers, as 'category A' agents, representing the most potent threats to national security.

David Walker, head of the new NIAID-funded regional biodefence centre in Galveston, argues that anthrax is the category A agent that is most easily prepared and disseminated by bioterrorists. And, as health workers saw during the anthrax mailings, antibiotics are ineffective against the bacterium after it begins to produce toxins in the body. "I think the emphasis on anthrax is

well-placed, because we've already faced it and we're going to face it again," Walker says.

But other researchers argue that a narrow focus on a small group of pathogens will prove counterproductive in the long run. "There is a lot more emphasis put on organisms on a list than is either desirable or necessary," claims Stanley Falkow, a microbiologist at Stanford University in California. He argues that diseases such as

influenza and other respiratory-tract infections routinely kill far more people than would die in a bioterrorist attack, and therefore deserve a greater share of the NIAID budget.

Distorted picture

If researchers were convinced that the increased emphasis on biodefence has been achieved entirely through 'new' funding that would otherwise never have found its way into NIAID's coffers, there would probably be few complaints. But incidents in recent months have raised fears that the biodefence push is distorting NIAID's overall priorities.

In July, for instance, the agency was forced to cut its grant awards for this year by \$117 million, to fund the development of an anthrax vaccine. AIDS researchers in particular were up in arms over the decision. John Moore, who works on HIV at Cornell

G. GAUGIER/SPL

CAMERA B. DOWSETT/SPL

M. ABBEY/SPL

EYE OF SCIENCE/SPL

University's Weill Medical College in New York, claims that NIAID has since withheld the 'bridging funds' that it normally awards to scientists whose research proposals do not quite pass muster on first review but are likely to be approved in the next round if they revise their grant applications in response to referees' comments. NIAID officials say that the budget for bridging funds is the same this year as last. But Moore's complaints are symptomatic of the distrust that has been sown by the diversion of funds to the anthrax-vaccine programme.

Given that his own background is in AIDS research, Fauci is sensitive to the concerns of those working on HIV. But he argues that the anthrax-vaccine situation is atypical. The US Congress is still trying to pass the 'BioShield' bill to fund the development and purchase of countermeasures against biological attacks. Once it is passed, Fauci says, there will be no need to raid NIAID's funds in this way. "It was an outlier," he says.

Fauci also argues that the increase in biodefence research, in particular that on

ways to boost human immune responses to bioterror agents, will spill over into other areas of infectious-disease medicine. "The commitment we have made to this area is extraordinary," he says. "It cannot help but boost the entire field." Fauci points in particular to an allocation of \$85 million over five years to establish five 'Cooperative Centers for Translational Research on Human Immunology and Biodefence'.

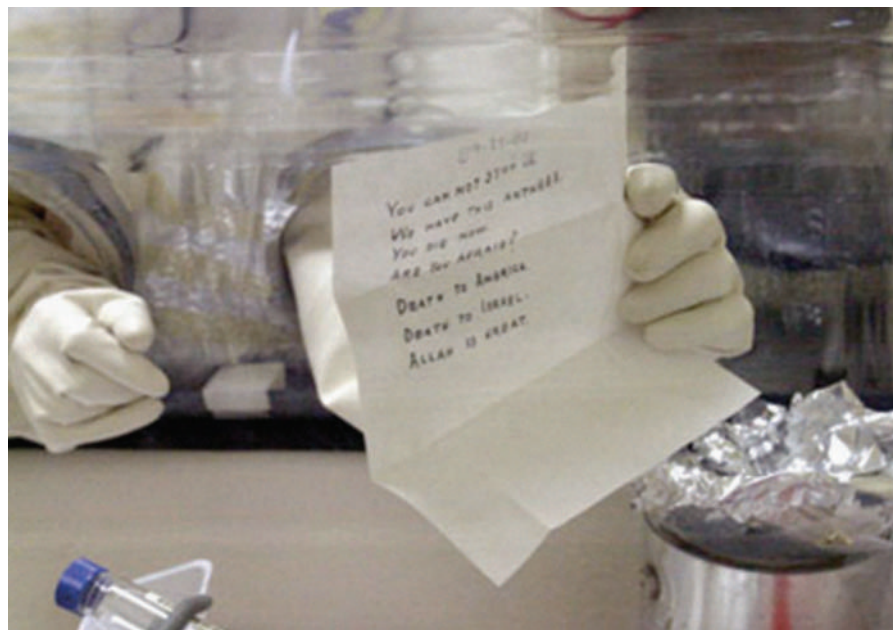
Act natural

Some of the work funded under this programme is directed at combating natural disease outbreaks rather than bioterrorist threats. For instance, Stanford microbiologists Ann Arvin and Harry Greenberg have received \$15 million to study influenza vaccines. They will compare two different types of vaccine: the traditional jab containing killed flu viruses, and an inhaled vaccine based on live, weakened viruses, that was approved by the US Food and Drug Administration this year. In studying immune responses to both types of vaccine,

Arvin and Greenberg will investigate such fundamental questions as how cells of the immune system communicate with one another.

Another new project, based at Emory University School of Medicine in Atlanta, Georgia, will compare the body's immune responses to anthrax and yellow fever vaccines. Rafi Ahmed, who heads the Emory Vaccine Center, says that the work will include investigations of how the vaccines induce long-term immunity. So far, this has been studied extensively only in lab animals. "I still think of this as basic research, but now it's basic research in humans instead of mice," Ahmed says. "It's a great opportunity to do some fundamental work on human immunity that is very relevant to the biodefence effort."

Behind such scientific questions, however, lurk a host of other issues — not least the risk of proliferation from the increased number of labs working on potential bioweapons agents. "I think our security will decrease, because access to dangerous pathogens and expertise in working with them is going to



Anthrax alert: the US battle against bioterror escalated in the wake of postal anthrax attacks in late 2001. Clockwise from left: a tainted letter sent to a US senator; members of the Coast Guard decontaminate after working in a Senate building; a ten-day ration of the antibiotic Cipro; a postal worker takes precautions while sorting mail.



increase vastly, and along with that will go a vast increase in the possibility of accidental escape, misuse, theft and bioterrorism," argues Barbara Rosenberg, an expert on bioweapons issues with the Federation of American Scientists in Washington DC. Her concern is underscored by evidence gathered by the FBI, in its inquiry into the 2001 anthrax attacks, suggesting that the perpetrator may have had links to the US military's biodefence establishment.

Given the risk of proliferation, the federal government has called on researchers to demonstrate that they will behave responsibly. In response, leading scientific journals have drafted policies to deal with biological research that has the potential for misuse¹. But many experts argue that the issue needs to be addressed further upstream, at the point at which research proposals are considered for funding. In October, an expert panel of the US National Research Council, chaired by Gerald Fink, a geneticist at the Massachusetts Institute of Technology, recommended that the federal Department of Health and Human Services set up a 'National Science Advisory Board for Biodefence'. This body would review all research funded by any US source, including private industry, that falls in one of seven "categories of concern", including work that would make a virus more deadly. As *Nature* went to press, researchers were still awaiting a response to these recommendations from the federal government.

Fear factor

In parallel, federal officials have been tightening the regulations that govern research on potential bioweapons agents, and stepping up their scrutiny of researchers who are working on such pathogens. Some researchers argue that this is creating a climate of fear that will drive talented individuals away from the field, however attractive the funding opportunities. "I'm afraid of these people," says one researcher working with category A agents, who asked to remain anonymous. "The FBI hasn't been able to come up with the real perpetrators of the anthrax attacks, and they're under a lot of pressure." As a result, he claims, federal officials are determined to make an example of anyone who is caught in breach of regulations that many researchers find confusing.

The Butler case represents an extreme example of the pressures that are being brought to bear on researchers working on potential bioweapons agents. But other scientists have been troubled by the public scrutiny that work on biodefence can attract (see 'Under the microscope', above right).

Nevertheless, a quick scan of US researchers now working in biodefence reveals that top scientists have not yet been scared away *en masse*. Indeed, their ranks include

Under the microscope

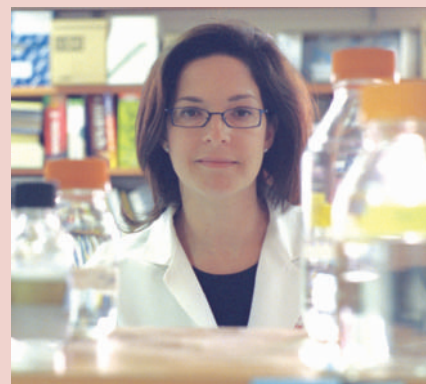
In May 2002, Ariella Rosengard of the University of Pennsylvania in Philadelphia published the results of a painstaking study that she hoped would help to defend against a future bioterrorist attack using smallpox. But she now wonders whether the publicity garnered by her project has actually undermined her research.

Rosengard was working with vaccinia, a relatively harmless cousin of the deadly smallpox virus. In her study, she mutated one of vaccinia's genes so that it produced a protein found in the smallpox virus, and then investigated how this protein inhibits enzymes that function in an arm of the immune system known as complement².

Clearly, anyone who repeated Rosengard's methods could make vaccinia into a much more dangerous virus. But she argued that the results opened avenues for developing drugs to counter a potential smallpox attack — and a commentary published alongside her paper agreed that the work was "far more likely to stimulate advances in vaccinology or viral therapy than it is to threaten biosecurity"³. In October, an expert panel of the US National Research Council, established to consider the proliferation risks posed by biological research, concluded that Rosengard's work will stimulate research into new treatments and vaccines.

Rosengard's paper attracted attention from the media, the scientific community and the US government. She was invited to meetings with Central Intelligence Agency officials and White House scientific advisers. She was also asked to speak at a meeting of scientists and government officials at the National Academies of Science about her decision to publish her work.

Rosengard didn't question her decision to publish until this spring, when she submitted a grant application to the National Institute of Allergy and Infectious Diseases (NIAID) in



Ariella Rosengard's research sparked fears that her methods could be adapted by bioterrorists.

Bethesda, Maryland. Given the direct relevance of her work to smallpox, Rosengard felt certain to get a slice of the biodefence funding pie. But it didn't even make the first cut. "I was surprised and perplexed — everyone was," Rosengard says. "Few others had any data to go on."

Rosengard's lack of funding may force her to shut down her lab. She is on leave from her post and has moved to Cambridge, UK, with her cardiologist husband. She is undecided over whether to apply for funding next year.

NIAID director Anthony Fauci says that the agency does not discuss individual grant applications. "What drives our awarding of grants and contracts is scientific merit — that always has been and always will be," he says.

But Rosengard cannot help but wonder whether the controversy over her paper influenced the decision not to support her work. "If you don't get funded, you have to ask whether people didn't understand what you were proposing, or whether there was some other reason," she says.



NIAID chief Anthony Fauci says biodefence funds will boost infectious-disease research in general.

some of the most highly regarded experts on infectious diseases. But even those who remain relatively relaxed about their regulatory environment are uneasy about the federal government's long-term commitment to the field. As a result, they say that they would think twice before recommending that young researchers commit themselves to specializing in biodefence projects. "I'm not sure there's enough

information out there yet for people to make informed choices about switching careers," says Clare Fraser, president of The Institute for Genomic Research in Rockville, Maryland. "We don't know how long this infusion of money is going to last, and I have yet to find anyone who is willing to make any guesses about how long that might be."

Some infectious-disease researchers, however, say that the new emphasis on biodefence has already caused a cultural shift in their discipline. Olaf Schneewind, who heads the NIAID-funded regional biodefence centre at the University of Chicago, argues that centres such as his have changed microbiology by bringing the ethos of 'big science' to the field. "The research groups have become larger, and the new centres involve a lot of different talents and efforts," he says. ■

Erika Check is *Nature's* Washington biomedical correspondent.

1. *Nature* 421, 771 (2003).
2. Rosengard, A. M., Liu, Y., Nie, Z. & Jimenez, R. *Proc. Natl Acad. Sci. USA* 99, 8808–8813 (2002).
3. Lachmann, P. J. *Proc. Natl Acad. Sci. USA* 99, 8461–8462 (2002).

QUAD system offers fair shares to all authors

Showing who did what could solve the problem of authorship abuse and disputes.

Sir— Authorship is the currency of modern science, the measure of one's contribution to the literature. Yet while the contents of an article are held to strict scientific standards, the recognition of authorship is far less so.

With no accepted system to guide the process by which authors are recognized for their individual contributions, and with the numbers of authors per article steadily increasing, the incidence of authorship disputes and abuse is rising, as Eugen Tarnow noted in Correspondence ("When extra authors get in on the act" *Nature* 398, 657; 1999).

Certain medical journals now require authorship declarations, although these are always qualitative. We suggest a quantitative method for evaluating authorship based on four categories of contribution: conception and design, data collection, data analysis and conclusions, and manuscript preparation. Each author would claim their percentage share of the total credit in each of the four categories. The least that one could contribute to a paper would be 10% within a single category, placing a theoretical limit of 40 on the total number of authors.

Authors would usually be listed in descending order of total contribution

across all four categories. When equal sums are reported, the listing could be made alphabetically, and for large collaborations, either groups or individuals may declare their contributions. Journals would be free to choose their own categories, but these would ideally be uniform across all papers.

The percentage share could be indicated as in the following example, reflecting a typical distribution in a student-mentor relationship in which the student, S, is responsible for 20% of conception and design, 90% of data collection, 80% of data analysis and conclusions and 20% of manuscript preparation:
 $S^{20,90,80,20}$, $M^{80,10,20,80}$.

The explicit nature of this system, dubbed Quantitative Uniform Authorship Declaration (QUAD), permits the reader to identify who contributed what, rapidly and easily. It should also discourage distortions of authorship by co-authors in positions of power and help reduce the numbers of honorary authorships.

Without external monitoring, it would be impossible to guarantee that all the authors have described their contributions accurately. However, the quantitative and transparent nature of QUAD should help reduce such behaviour.

Journals that have introduced

authorship declarations have seen only partial adoption by authors, but we hope that many more journals and authors will welcome the advantages offered by QUAD. Ultimately, this system will prove more useful than existing *ad hoc* methods when scientists and administrators are faced with funding and hiring decisions.

Justus V. Verhagen^{50,30,30,30*},

Karin J. Wallace^{30,30,10,†},

Stephan C. Collins^{20,40,20,10‡},

Thomas R. Scott^{0,0,20,50§}

*Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, UK

†Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK

‡The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK

§College of Sciences, San Diego State University, 5500 Campanile Drive, San Diego, California 92182-1010, USA

Nature's authorship policy is that authors are strongly encouraged to include a statement in the Acknowledgements to specify the actual contribution of each co-author. See www.nature.com/nature/submit/policies/index.html#2 — Editor, Correspondence

Tidewater glaciers move at their own pace

Sir— Arctic warming is receiving much scientific scrutiny and public interest, and your News Feature about Alaska's climate "Too hot to handle" (*Nature* 425, 338–339; 2003) gives a generally accurate overview of the reasons why the Arctic deserves our continued attention. However, one ongoing event in your list of direct consequences of recent Arctic warming, the rapid retreat of Columbia Glacier, should not be there.

Columbia Glacier is one of many tidewater glaciers — glaciers that terminate in the ocean but do not float — that fringe Alaska's central and southern coasts. They are well-known to advance and retreat out of synchrony with land-terminating glaciers, and even with each other. Climate appears to control tidewater glacier advance and retreat only on long (millennial) timescales; shorter annual, decadal, and even century-to-century comparisons do not show a close relationship between tidewater glacier activity and climate.

Early European explorations of Alaska at the end of the eighteenth century showed that tidewater glaciers in southern Alaska were fully extended, with termini reaching all the way to the coast, but during the nineteenth and early twentieth centuries these glaciers rapidly retreated, exposing such long and deep fjords as Johns Hopkins Inlet and Disenchantment Bay. Columbia Glacier attained its advanced position between the fifteenth and nineteenth centuries, and only since the early 1980s has it started its retreat.

The end of Columbia Glacier has retreated roughly 13 kilometres from its pre-1980 position, and continues to retreat at roughly half a kilometre per year while discharging some 11 cubic kilometres of ice per year into Prince William Sound. But as Columbia Glacier retreats, Hubbard Glacier, 400 kilometres to the southeast in Disenchantment Bay, has been advancing since 1980.

None of this is to say that tidewater glacier behaviour is irrelevant to present-day concerns about climate change. Tidewater glaciers are affected by climate change on long timescales (primarily through accumulation of mass by snowfall

and loss by melt), but are even more strongly affected on short timescales by the ocean, through the influence of heat advection and forces of hydrostatic water pressure. These factors add great complexity to tidewater glacier dynamics and result in nonlinear and irreversible responses to climate, which on long time-scales does indeed control all glacier growth and contraction.

Arctic warming is very real, but the complicated interaction between tidewater glaciers, oceans and climate — and the likelihood that tidewater glacier retreat may not be caused directly by the current climate — should be more widely recognized.

W. T. Pfeffer

Department of Civil, Environmental and Architectural Engineering and Institute of Arctic and Alpine Research, Campus Box 450, University of Colorado, Boulder, Colorado 80309-0450, USA

correspondence

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter. Published contributions are edited.

Playing God?

Scientists must let society decide on the ethics of reproductive technologies.

Whose View of Life? Embryos, Cloning and Stem Cells

by Jane Maienschein

Harvard University Press: 2003. 304 pp.

\$27.95, £18.50

Robert Winston

In 1692, Nicolaas Hartsoeker drew what he thought he saw through his microscope: a human sperm, containing in its head a homunculus, a perfectly formed little person. This arresting view of life's beginnings still reminds us that scientists are not always objective; they can be persuaded to see what they believe to exist. The preformationists, who believed that sperm contain miniature people, followed a line of thought stretching back to Aristotle. A hundred years later, Rabbi Pinhas Elijah recalled Hartsoeker's drawing and argued that the ancient sages were right after all — destruction of the seed is murder, as it destroys little persons, and masturbation is wrong.

Pinhas Elijah's reasoning was impeccable, as far as it went. But his ethics were flawed because we now know that the head of the sperm does not contain a person. Our ethical understanding must depend on accurate observation of the natural world; consequently, ethical positions must change as human knowledge improves.

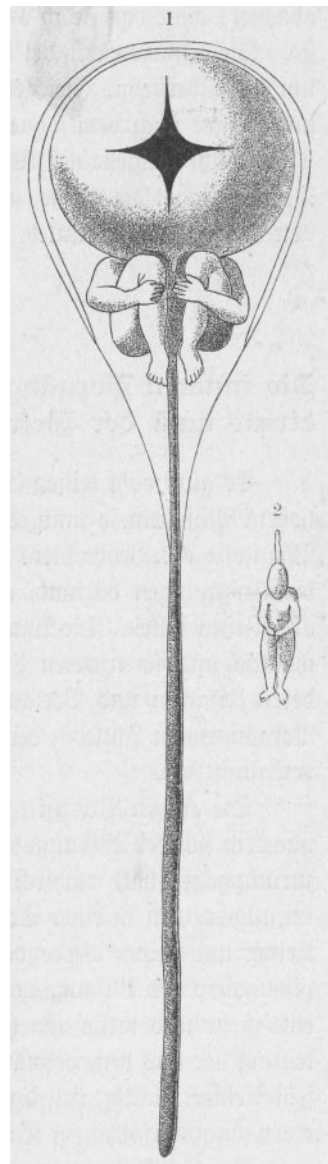
Jane Maienschein's latest book, *Whose View of Life?*, starts with an account of the preformationists. She maintains that the ethical, political and social issues surrounding modern reproductive technologies can be better understood with a historical perspective. She traces the history of reproduction from the recognition by Karl Ernst, Ritter von Baer, that mammals produce eggs, through Ernst Haeckel's notions of comparative embryology, to Thomas Hunt Morgan's early understanding of genetics, and modern *in vitro* fertilization (IVF) and stem-cell biology.

This is an extremely valuable approach. Scientists have repeatedly made errors of judgement through lack of objectivity, as Hartsoeker did. They have also sometimes arrogantly assumed that they know what is best for society — the pursuit of eugenics in the United States, Britain and Germany being a case in point. Maienschein cites some chilling modern examples; for instance, in the debates over cloning, modern scientists often ignore the cultural sensitivities of their society. She emphasizes that if science is to improve the lot of humans, then scientists must hone their moral thinking and be more ready to listen to the opinions of ordinary people.

Maienschein's historical account is both engaging and accurate, but there are certain gaps in this excellent book. Much of the book is centred on experience in the United States; there is no reference to the growing interest in these technologies in the Far East, and no mention of religious attitudes outside the Judaeo-Christian tradition. Nor is there any analysis of the debates in either Britain or Australia — two countries that have contributed extensively to advances in this field. After all, the first IVF babies were British, as was the first mammalian clone, and early work on embryonic stem cells was initiated in Britain too.

There is no reference either to the deliberations of the Warnock Commission, nor much mention of British or European parliamentary debates about IVF or stem cells. And Maienschein ignores the fact that many people consider proper regulation to be the best way of controlling these technologies. Europeans are concerned at the lack of regulation in the United States, a situation so different from the European and Australian models that these should surely be worthy of some consideration.

A more thorough discussion of religious attitudes would also help. Western ethics centre on a belief in the sanctity of human life, and religious principles guide much deliberation about these technologies. In the United States and Europe, many people profess some belief in God, so it would help to understand some of the religious views. For example, it is an oversimplification to imply that orthodox Jews believe that ensoulment occurs 40 days after conception. And Maienschein ignores the



The little people? Drawings by Nicolaas Hartsoeker and Dalepadius (inset) show their preformationist views of sperm.

paradox that “playing God” is by no means necessarily evil — many religious people hold that *imitatio Dei* is morally positive. If humans are indeed made in the image of God, then playing God — using God-given intelligence and hence technology — to promote and enhance life may be highly ethical.

Better discussion of some ethical principles would also be relevant. Why the “horror” of human cloning? If cloning could be developed without the risk of producing abnormal offspring, what would be the objection then? Is this an affront to human dignity, given that around 1 in 250 human births — identical twins — are already perfect clones?

Maienschein correctly points out that promoting an understanding of science is the first step in wise biopolicy decision-making, and she helpfully suggests how this might be achieved. But more knowledge may bring more public mistrust, not less. Unless scientists accept that they do not own the science, and are not masters of technology, this mistrust is likely to grow. There are massive mountains to climb. A difficult change is needed before scientists accept that society must

decide what it feels is in its best interests. Scientists need to show more objectivity, better ethical values and more concern about the commercial aspects of their work.

The US-centric position that Maienschein adopts is disappointing. It fails to appreciate the growing global anxiety that big business, often based in the United States, will control and selfishly exploit the fruits of science. Is this likely to benefit most of us? ■

Robert Winston is at the Institute of Reproductive and Developmental Biology, Hammersmith Hospital, Imperial College, London W12 0HS, UK.

The two faces of consciousness

Consciousness: An Introduction

by Susan Blackmore

Hodder and Stoughton: 2003. 544 pp. £19.99

Oxford University Press: 2003. 460 pp.

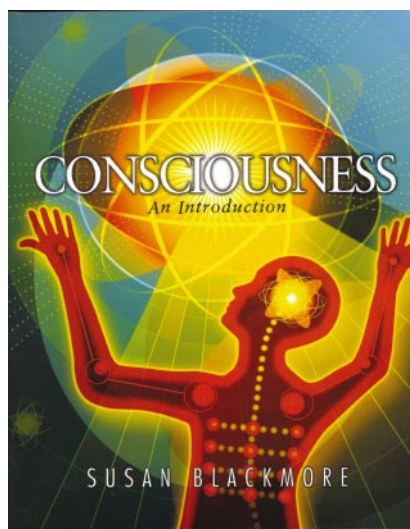
\$39.95 (pbk), \$79 (hbk)

Ilya Farber

The study of consciousness being rather broad, its more science-oriented practitioners quickly become adept at glancing through promising-looking new books to discern whether they partake a little too much of the 'cosmic'. This isn't usually difficult, but Susan Blackmore's new book almost seems to have been intentionally designed to thwart such attempts at classification. *Consciousness: An Introduction* is clearly intended and marketed as a college-level textbook, but it sports a cover that is dominated by a glowing humanoid outline festooned with chakra-like circles and reaching ecstatically towards some sort of celestial radiance.

A glance through the copious figures initially reveals a reassuring profusion of brains and graphs and experimental protocols, but in the final quarter of the book these give way to buddhas, drug-induced visions and floating spirits. Even the usually reliable strategy of investigating the author's credentials just generates new puzzles. A search through Blackmore's oeuvre turns up near-death experiences and memes, a book on testing your psychic powers and a chapter on why parapsychology tells us nothing about consciousness. For decades she described herself as a hopeful and open-minded investigator of psychic phenomena, but was all the while on the board of the magazine *Skeptical Inquirer*, and in 2001 she announced the end of her involvement in such studies. It is easy to imagine potential readers or teachers giving up in bafflement, unable to figure out just what kind of book this is.

That would be a shame, though, because Blackmore has in fact written a truly excellent textbook. Two years ago she gave up her university position to work on this project full-time, and the attention shows. The book's 27 detailed chapters cover every major aspect of consciousness, and Blackmore reveals herself to be a careful and judicious researcher: the central figures and theories of most chapters are familiar, but they are presented with unusually careful attention to details of argumentation and evidence, rather than being reduced to the sort of simple iconic positions that fuel most 'battle of ideas' texts. Many lesser-known authors and positions are also introduced, and some of the most interesting ones receive extended treatment. The book is nearly encyclopaedic



Books about consciousness divide into the scientific and the 'cosmic' — but which is this?

in its comprehensiveness, but the discussion is carefully structured: ideas appear not as individual items to be memorized, but as steps in an integrated, multifaceted process of investigation.

The book does have a bit of a split personality, seeming at times like a guidebook to self-discovery and at others like a primer on recondite academic disputes. It soon becomes clear, however, that this bipolar character is strategic. The chattier bits (full of second-person questions and exhortations) provide excitement and motivation, and encourage students to adopt an engaged, reflective approach to the material, whereas the more scholarly parts delve unapologetically and at length into the details of particular theories and problems. This double structure should make the book accessible and attractive to a wide range of readers.

There are a few notable problems. One is an unfortunate tendency for favourite authors, such as Daniel Dennett and Benjamin Libet, to crop up again and again in different contexts, sometimes in place of others who would arguably be more relevant. Blackmore draws on a wide variety of philosophically engaged scientists (including Francis Crick, Antonio Damasio, Ray Jackendoff, Vilayanur Ramachandran and Francisco Varela), but is much less careful with scientifically engaged philosophers, often letting Dennett stand in for the whole pack. This is most problematic in the chapters on neuroscience, where one would have liked to see more coverage of the specific proposals of philosophers who specialize in the topic, such as Patricia and Paul Churchland, Kathleen Akins, Ned Block, Owen Flanagan and Thomas Metzinger. Blackmore, an experienced author of popular-science books, often uses the journalist's technique of using individual scholars as symbolic representatives

of broader positions, a practice that is less appropriate in the academic context and only exacerbates the above problems. The later chapters on parapsychology, hallucinatory states and meditative spiritual traditions are also less successful than the rest; Blackmore's combination of expertise and scepticism makes her an ideal 'tour guide' for such realms, but ultimately these chapters establish little and do not seem to connect with the ideas developed in the earlier sections.

Ultimately, this remains a very satisfying book. It could serve well as a core text for courses in philosophy, psychology and related disciplines, and would provide useful context for other, more discipline-specific texts. Its broad scope and clear explanations also make it an excellent choice for independent study. It's a shame about the cover, though. This might seem a petty complaint, but considering that academic respect for the study of consciousness is still grudging and probational, such matters of appearance are not trivial. I'll be using *Consciousness* in my course next semester, but I'll also be passing out a nice selection of book covers on the first day of class. ■

Ilya Farber is in the Department of Philosophy, George Washington University, Washington DC 20052, USA

Rich rewards

Pioneers of Microbiology and the Nobel Prize

by Ulf Lagerkvist

World Scientific: 2003. 182 pp. \$24, £16

W. F. Bynum

Historians of twentieth-century science often write about their subject in terms of Nobel milestones. Laureates are convenient markers of major scientific advance, and have the added attraction that a great deal of information about them is generally in the public domain. This in itself reinforces their scientific visibility and assures us that, on the whole, the various bodies that Alfred Nobel asked to adjudicate his prizes have done their job correctly. Disciplines such as mathematics, evolutionary biology and geology, overlooked by Nobel, have their own means of rewarding excellence, and historians use these accordingly.

That scientists also value the award of the ultimate scientific prize far beyond its considerable monetary value has recently been demonstrated by the contested 2003 prize for medicine or physiology, which was awarded for medical magnetic resonance imaging (see *Nature* 425, 648; 2003). Although it took some time for the current prestige of the Nobel prizes to develop, the second award in physiology or medicine (1902) exacerbated a priority dispute between

Winging it

The colour and intricate detail of a moth's wing, like the one shown here from the painted lichen moth (*Hypoprepia fucosa*), is rarely appreciated. Although there are many more species of moths than butterflies, their nocturnal habit tends to keep them from view. Print-maker Joseph Scheer became interested in moths when he began using a high-resolution scanner to capture directly images of insects. The prints allow a view of the anatomical detail, previously only available through a stereomicroscope, across an entire specimen. Now Scheer has become something of an amateur lepidopterist, studying the diversity of the local moth populations as he collects more specimens for his prints, a selection of which can be seen in *Night Visions: The Secret Designs of Moths* (Prestel, £29.95).

Mary Purton



Ronald Ross (who got the prize) and Giovanni Battista Grassi (who was overlooked) over their relative contributions to knowledge of the role of *Anopheles* mosquitoes in malaria transmission. Nobel had made provision for each of his prizes to be shared by up to three people, although shared prizes in science were rare before the inter-war period and became the norm only after the Second World War.

Ulf Lagerkvist's discussion of the early medicine or physiology prizes is by far the most original part of his little book, which focuses on the lives of four pioneers of microbiology who all won the coveted prize: Emil von Behring, Robert Koch, Paul Ehrlich and Elie Metchnikoff. He describes the composition and deliberations of the early committees of the Karolinska Institute, the institution charged by Nobel with choosing the laureate in that loosely defined area. Johan

Erik Johannson, professor of physiology at the Karolinska, probably persuaded Nobel to add 'physiology' to the medical prize. Johannson campaigned ardently against the award of a prize to Paul Ehrlich, his reason being that Ehrlich's famous 'side-chain' theory of antigen-antibody interaction was too speculative and had been contested by Svante Arrhenius, the deserving winner of the 1903 prize for chemistry.

Grumbings about nationalistic bias surfaced early in international attitudes towards Nobel committees, and it must be said that the first two Scandinavian medical laureates, Niels Finsen (1903) and Allvar Gullstrand (1911), have not exactly remained household names, if they ever were. Johannson may have been unusual in raising the Scandinavian issue, and the internationalism of the early prizes is evident in all the sciences. The medical committee was also adventurous

in how it shared a couple of early prizes. Two of Lagerkvist's four principals — Ehrlich and Metchnikoff — split the 1908 prize, and the committee had already divided the 1906 prize between Camillo Golgi and Santiago Ramón y Cajal. In both cases, the recipients held diametrically opposed views on their research areas, Golgi and Ramón y Cajal publicly disputing the structure and function of neuronal elements. Ehrlich and Metchnikoff represented chemical and cellular approaches to the main determinants of immunity. In contrast, the early shared physics prizes went to people who had worked together, and the chemistry committee did not award a shared prize until 1929.

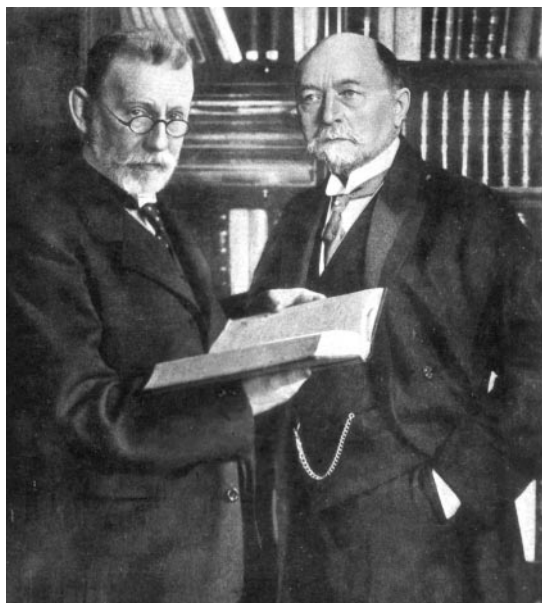
The other awkwardness that had to be resolved before

the first awards was Nobel's stipulation that the science prizes ought to be awarded for the best research 'in the preceding year'. Members of the awarding bodies knew that it takes time for the permanent value of a scientific discovery to be judged, and one of Nobel's original trustees, Ragnar Sohlman, managed to get the terms of the will relaxed. The Medicine or Physiology Committee seemed to have adopted a rough ten-year guideline, which explains why some of the early recipients were deemed eligible. Thus the first winner, von Behring, had done his important work on diphtheria (the word is irritatingly misspelled throughout Lagerkvist's book) roughly a decade earlier. By the time of the award he was engaged in largely fruitless research on tuberculosis.

Few would quibble with the decision to crown each of Lagerkvist's four pioneers. From the historical perspective, however, even the informal ten-year rule of the Committee creates problems. Koch was a giant of bacteriology in the 1870s and 1880s, but his later years, with tuberculin, the radical distinctiveness of bovine and human tuberculosis, and the wanderings in Africa investigating tropical diseases, did not produce the stuff of greatness. By 1908, when Metchnikoff received his prize, he was more obsessed with intestinal hygiene and longevity than with basic immunology. Only Ehrlich remained unremittingly dedicated to his laboratory craft.

Some of these issues have to be teased out of Lagerkvist's volume, which tries to do too much in too little space. The first half, a potted biographical history of medicine, has been done many times before, and should have been omitted in favour of a more extensive and subtle examination of the achievements and careers of his four principals, each of whom left permanent legacies. ■

W. F. Bynum is at the Wellcome Trust Centre for the History of Medicine, University College London, 24 Eversholt Street, London NW1 1AD, UK.



Nobel prizewinners: Paul Ehrlich (left) and Emil von Behring.

AP

To put it simply...

Massimo Piattelli-Palmarini

Try a simple experiment. Stand beside a window, take a look at the scene outside and then sit down and quickly write a half-page report of what you have seen. It is a safe bet that you will mention people walking, cars, buses, streets and buildings, or (depending on the chosen window) grass, trees, hills, rivers and birds flying. It's extremely unlikely the report will mention blue sport utility vehicles, belted kingfishers, or instances of strutting, ambling or swaggering — even if your automotive, avian or human-locomotion expertise allows you to identify things and events at that level of detail. At the other extreme, it did not cross your mind to write expressions such as “intentionally cause their body to move horizontally”, or “self-propelled wheeled vehicle”, which some philosophers might use. Interestingly, your report will contain only terms that are the first to be learned by a child, are usually expressed by a single word in most languages, are remembered best and are preferentially used when we ‘talk to ourselves’. They are, in the cognitive-sciences jargon, ‘basic concepts’.

Ideally, in a theory of concepts and in lexical semantics, the term ‘basic’ ought to cover the topmost level of abstraction (something like the undifferentiated essential furniture of the world). Or, at the other extreme — in a tradition that extends back to the philosopher David Hume — they have roots in our most direct access to unadorned sense impressions (a green splotch here and now in front of me). Unfortunately, ‘basic concepts’ sit comfortably at an intermediate level — they are neither too general nor

too specific. Lunch is a basic concept, but so are bread, spoon and banana, all possible parts of that lunch. The desire to decompose such concepts into real basics has proved to be almost irresistible. A famous example is the (alleged) decomposition of ‘kill’ as ‘to cause to become not alive’. It is claimed that in our mental lexicon, kill is just shorthand for the latter composite expression. But consider situations in which the cause is separated from its outcome by a long and/or anomalous chain of intervening events, for example, the suicide of a government advisor after his involvement in a media row had been made public the previous week. It is clear that we will consider this an instance of ‘causing to become not alive’, but not an instance of ‘killing’. Therefore, it seems that if any such decomposition is to hold water, we need an additional component X, so that: kill = CAUSE + BECOME NOT ALIVE + X. That would be acceptable if X were both general (in the same league as CAUSE), and sufficient. But it turns out that X needs to be as specific as ‘kill’. So there is no gain in understanding — and therefore no explanatory use — for any such decomposition.

It has been widely assumed in the philosophy of mind, psychology and lexical semantics that basic concepts must be homogeneous under some interesting description or other. For example, they may have a characteristic role to play in concept acquisition, perception, memory or thought. However, there is no serious evidence for any such claim, and when combined with a thorough account of concept possession, this disconcerting fact appears to lead to implications that are utterly implausible. For instance, it suggests that similarities among things imply similarities among concepts of those things. A rope may be similar to a snake because of their shape, but our concept of a rope is not at all similar to our concept of a snake. A leading thinker in this domain, Jerry Fodor, concludes that basic concepts have nothing interesting in common except their basicness — basic concepts are boring. But, he hastens to add, the fact that they are isn't!

Fodor champions a different tack altogether, treating basic concepts as ‘atoms’ that cannot be decomposed, expressing exactly the property that they express (for example, the property of killing) — no less and no more. There are two core components of this atomistic approach: first, a causal link between our mind and the property being exemplified, or evoked in discourse; and second, some efficient way of presenting a good example of that very property. Better still would be a prototypical

Basic concepts

Why has cognitive science struggled to find an explanation for the terms that we use every day?

example, as no one would introduce a child to the concept ‘bird’ by displaying a penguin, or to the concept ‘tree’ by displaying a bonsai. There is every reason to suppose that the human mind is natively equipped with the capacity to lock onto the salient property after an encounter, as long as it is supplied with the correct mode of presentation and the appropriate situation. The capacity to generalize instantly and competently from a good example, while retaining the corresponding verbal label, remains awesome, and at present mainly mysterious.

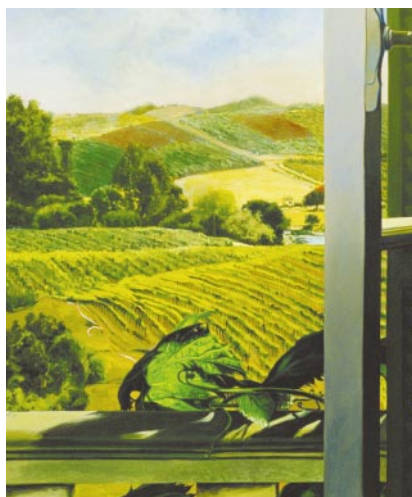
It is tempting to make direct connections between the meaning of a concept and the most obvious inferences that the person who possesses it is disposed to make — for example, the inference going from ‘bird’ to ‘animal’ and from ‘water’ to ‘wet’. This move goes under the name of inferential role semantics. The counter to this move is that the content of a concept cannot ultimately consist of any kind of readiness to do something, not even a ‘disposition’ to draw inferences. It is rather a mental particular that applies to things, notably to the standard instances of the category for which the concept stands. But even failure to identify marginal exemplars (is a whale a fish? Is mercury a metal?) does not count as failure to possess that concept fully.

As much of our mental life consists of applying concepts to things, it may come as a disappointment that the most plausible attempts at an explanation (decomposition, inferential power, strategies of verification) did not work. Halfway between an avowal of impotence and an incitement to do better, the theory of concepts seems to be a signal place where, in spite of a voluminous literature, cognitive science “went wrong” — at least until now.

Massimo Piattelli-Palmarini is in the Department of Linguistics, University of Arizona, 1103 East University Boulevard, Tucson, Arizona 85721-0025, USA.

FURTHER READING

Fodor, J. A. *Concepts: Where Cognitive Science Went Wrong* (Oxford Univ. Press, Oxford, 1998).
Margolis, E. & Laurence, S. (eds) *Concepts: Core Readings* (MIT Press, Cambridge, Massachusetts, 1999).
Murphy, G. L. *The Big Book of Concepts* (MIT Press, Cambridge, Massachusetts, 2002).



Taking in the view: we tend to use simple language when asked to describe a scene.

Epidemics-in-waiting

Jim Bull and Dan Dykhuizen

Could the next SARS-like virus reach epidemic proportions? Quantifying the likely threat of emerging diseases isn't easy, but evolution is a crucial factor that may tip the balance in favour of such human parasites.

One of the oldest tenets of evolutionary biology is that it is easier to change a little than a lot. We also know that evolutionary change is more easily selected for in a large population than in a small one. On page 658 of this issue, Antia *et al.*¹ combine these facts to reach a previously unappreciated conclusion about emerging infectious diseases: some types of infectious parasites that attack the human population may pose a serious threat even if they are not initially able to cause epidemics. The reason is that certain parasites are specially poised to evolve so that they can cause epidemics.

In epidemiological models, an infectious agent can be characterized by its basic reproductive rate, R_0 . This is the average number of new infections caused by the first infected individual in the population. The epidemic threshold is $R_0 = 1$, above which the disease spreads (neglecting random effects) and below which it eventually dies out. R_0 is not, however, a measure of virulence or of the harm inflicted on the host by infection. In humans, for example, R_0 is virtually zero for some diseases with a high mortality rate (such as rabies, and hantavirus respiratory infections); but R_0 is well above unity for other diseases that also have high mortality rates (such as measles in undernourished populations, AIDS and smallpox).

Antia *et al.*¹ point out that, because of evolution, parasites with an R_0 value that is hovering just below one can be epidemics-in-waiting. An obvious reason for this is that it takes less change to achieve an R_0 of more than one if the initial R_0 is just below one. A less obvious reason — and the focus of Antia and colleagues' paper — is that the length of time a parasite persists in the population before disappearing increases with R_0 ; parasites with R_0 nearly at unity can persist for a considerable time by chance. The longer the parasite persists, the greater will be its opportunity to evolve to a higher R_0 . This is merely a population size effect: each additional host infected before the parasite dies out provides yet another opportunity for a mutation that might push R_0 over the epidemic threshold.

The model is tractable because of its reliance on the single parameter R_0 , but the biology behind it is potentially complicated.



Figure 1 **Danger ahead?** The virus (inset) causing severe acute respiratory syndrome, or SARS, emerged in southeast Asia in late 2002. As the infection spread, face masks became a common sight on the streets of Hong Kong and elsewhere. Antia *et al.*¹ describe a way to help identify other viruses with epidemic potential.

R_0 encapsulates the long chain of events involved in the parasite's association with its host — its first contact, entry, growth in the initial tissue infected, infection of secondary tissues, and so on, until the final stage of its dissemination to make contact with other hosts. A mutation that increases R_0 may arise at any stage in this chain, provided that it ultimately leads to an increase in the number of infections. For instance, an increase in R_0 may evolve through changes in surface molecules that improve parasite infection of new hosts directly, or it may evolve through improved growth within the host (possibly increasing virulence) so that more parasite progeny are disseminated from the host, resulting in higher numbers of secondary infections.

The rabies and hantavirus examples mentioned above are characterized by good within-host growth of the parasite but poor dissemination to new hosts. The 1976 outbreak of Ebola virus in southern Sudan (which spread from an initial infection of a cotton factory worker to the owner of the local jazz club, then to others at the club, giving at least eight generations of transmission²) is a case in which within-host growth

was good and the dissemination–infection stage brought the parasite perilously close to the epidemic threshold. However, a parasite could also begin its foray into humans by being fairly infectious but poor at surviving the onslaught of our immune system (as seems to have been the case with the Ebola virus that, in 1989, destroyed an imported Philippine monkey colony in Reston, Virginia³). Mutations that increase R_0 might then improve within-human growth. This type of emerging pathogen is the most easily missed and potentially the most dangerous. Efforts to understand the relationship between parasite adaptation to hosts, virulence and transmission have developed into a small industry in evolutionary biology. The relationship is complicated because it involves group versus individual selection, population bottlenecks and trade-offs^{4,5}.

Antia and colleagues' result¹ contributes to a growing awareness of the evolutionary and ecological factors surrounding the emergence of new diseases. We have already had warnings that virulence itself may evolve in response to changes in cultural practices⁶, and that immunocompromised patients

P. PARKS/AFP

CAM/RA. B. DOWSETT/SPL

may act as stepping stones to foster evolution of new pathogens capable of attacking people with healthy immune systems⁷. With the possibility of using grafts from non-human species to replace tissues in humans, we also need to be aware of the potential for activation and genetic recombination of otherwise dormant retroviruses in the human genome or in the graft.

Antia *et al.* do not, however, emphasize cultural factors in disease emergence. Instead, they provide a way of identifying which agents are most worthy of attention — those closest to the epidemic threshold. An example suggested by Antia *et al.* is the threat of monkeypox in a world with little resistance to its relative smallpox, because of a lack of either vaccination or exposure. Furthermore, the result draws attention to the neglected topic of parasite dynamics in the pre-epidemic stages. This was brought into focus earlier this year when it was realized that the initial spread of severe acute respiratory syndrome (SARS; Fig. 1) depended heavily on the social connectivity of the first (index) case in a community. Such results and realizations give us a better understanding of how to contain infectious diseases, through early prevention rather than cure. Ultimately, we should learn where and when to apply our efforts to block transmission and so prevent an epidemic.

Currently, the resources and public attention devoted to an infectious disease depend on a combination of social, biological, economic and political factors specific to that disease. Disease virulence, transmissibility and incidence are included in such considerations. The complacency of the pre-HIV and pre-bioterrorism eras has yielded to a growing acceptance of the need to monitor pathogens and even pre-pathogens in our environment. It is not beyond imagination that, even with existing technology, methods could be developed for monitoring emerging pathogens, potentially distinguishing between strains with differing R_0 values. The means, provided by Antia *et al.*¹, of identifying these epidemics-in-waiting could become a critical tool in a global defence strategy against emerging pathogens. ■

Jim Bull is in the Section of Integrative Biology and the Institute of Cellular and Molecular Biology, University of Texas, Austin, Texas 78712, USA.
e-mail: bull@bull.biosci.utexas.edu

Dan Dykhuizen is in the Department of Ecology and Evolution, State University of New York at Stony Brook, Stony Brook, New York 11794, USA.
e-mail: dandyk@life.bio.sunysb.edu

1. Antia, R., Regoes, R. R., Koella, J. C. & Bergstrom, C. T. *Nature* **426**, 658–661 (2003).
2. Garrett, L. *The Coming Plague* (Farrar, Straus & Giroux, New York, 1994).
3. Preston, R. *The Hot Zone* (Anchor Books, New York, 1994).
4. Frank, S. A. *Q. Rev. Biol.* **71**, 37–78 (1996).
5. Frank, S. A. *Immunology and Evolution of Infectious Disease* (Princeton Univ. Press, 2002).
6. Ewald, P. W. *Evolution of Infectious Disease* (Oxford Univ. Press, 1994).
7. Wallace, B. *Am. Nat.* **133**, 578–579 (1989).

Quantum optics

Light at a standstill

Marlan O. Scully

Nothing travels faster than light, but how slow can light go? Pulses of light have already been slowed to speeds of just a few metres per second, but now they have been brought to a complete halt.

Frozen light is a reality. As they report on page 638 of this issue, Bajcsy and colleagues¹ have trapped, and held, a pulse of light for a few hundredths of a millisecond, which is a long time in optical terms. To those unfamiliar with the realm of quantum optics, the notion of stationary light may seem rather strange. But ultraslow light — travelling at just a few, instead of 300 million, metres per second — is easily available in many labs these days. So bringing a light pulse to a full stop was the next logical step.

And it is of interest for many reasons. This is the latest development in a continuing paradigm shift in optics, occasioned by the marriage of quantum coherence in atoms and molecules with coherent light^{2–5}. The ability to trap and hold light also holds promise for application to such diverse areas as quantum informatics, nonlinear optics and even the foundations of quantum mechanics. Perhaps most important of all, it is simply fascinating science.

To put the achievement of Bajcsy *et al.*¹ into perspective, let me briefly mention some of the key steps and issues that have led to the creation of frozen light pulses. The landmark slow-light experiment⁶ — down to a speed of just 17 metres per second — was carried out in an ultracold gas. The advantage of such low temperatures is that they eliminate the spread in atomic frequencies that is caused by atomic motion: this optical Doppler effect could otherwise obscure, even spoil, the measurements of slow light. It is not necessary, however, to use trapped ultracold atoms and get rid of the Doppler effect to slow light down: this has been proved by the slowing of light to less than 100 metres per second in a hot gas of rubidium atoms⁷.

It could be argued that the know-how to make slow light has been around since the late nineteenth century. The velocity of the peak of an optical pulse is called the group velocity because it takes a 'group' of two or more frequencies of light to make a pulse. Atoms and molecules tend to interact with light more strongly at certain frequencies, and this is manifest in a variation of the material's index of refraction, which causes different waves to move with different velocities. Of course, this affects the group velocity of the pulse — and is the reason that light is dispersed by a prism or water droplets to make a rainbow.

The good news is that strong dispersion is

easy to achieve: the light frequency should be tuned close to the resonant frequency of the atoms, and then the group velocity may be slowed by a factor of a thousand or more. The bad news is that minimum pulse velocity also means maximum absorption of the light by the material: the pulses move very slowly, but they are so strongly absorbed that they penetrate only a few wavelengths' distance through the material.

But the last decade of the twentieth century saw revolutionary changes in optical science. We now have 'phase-coherent' materials, which give the huge dispersion associated with resonance but do not absorb light. These materials are truly a new state of matter, aptly called 'phaseonium', and their electromagnetically induced transparency⁸ has proved useful in slowing light down to just a few metres per second. It has been shown^{9,10} that the quantum state of a light beam can be stored in phaseonium vapour and then retrieved by a later pulse of light. This works because in slow-light propagation, part of the light beam is converted into atomic excitation (that is, it is mapped onto atomic spins). This transfer of the information contained in the light's quantum state onto the atoms (writing), together with its subsequent retrieval (reading), is a vital tool in the emerging field of quantum information, including cryptography and computing. However, although the concept of storing light by 'drawing' its quantum blueprint in the atomic medium is interesting, this is not the same as freezing light.

In fact, there are several possible routes to frozen light. For example, my own group is investigating the passage of slow-light pulses through a moving phase-coherent medium, such that the pulse would appear stationary to a fixed camera in the laboratory¹¹. The route chosen by Bajcsy *et al.*¹, following the theoretical outline of ref. 12, does not involve a moving medium. Instead, a light pulse is trapped inside a gaseous, phase-coherent medium by changing the optical properties of the medium. There are two steps to the process: first, two laser pulses, the 'control' and 'signal' fields, are injected into a medium in which all the atoms are in the lowest-energy, ground state; then, with both control and signal fields in the medium, the control pulse is turned off. This results in the atoms' absorbing the signal field and storing the quantum state of the signal. Following the example of previous work, the next step

would be to regenerate the signal field (or its time-reversed partner) using a forward-propagating (or backward-propagating) 'read' field in the same direction as the first coupling laser.

But in their experiment, Bajcsy *et al.*¹ put in forwards and backwards read fields simultaneously to regenerate the signal. Now the plot thickens. The forwards and backwards fields interfere, producing a spatially periodic variation of light intensity, which in turn results in spatially varying atomic absorption. This acts like a stack of mirrors, which reflect and contain the regenerated field. The two laser pulses play a dual role: they convert the stored state into a signal pulse, and modulate the atomic absorption to localize or freeze the regenerated field.

Frozen light naturally suggests several interesting applications. First, as there is little or no loss, there will be little or no noise. This augurs well for applying such techniques to quantum informatics, where the suppression of noise is vital to avoid decoherence. Furthermore, the nonlinear coefficients associated with phase-coherent matter are very large: in a stationary pulse, even tiny

amounts of light will cause large nonlinear effects, which have long been sought in both classical and quantum signal processing. One thing is clear — frozen, stationary pulses of light mark a new chapter in quantum optics.

Marlan O. Scully is at Princeton University, Princeton, New Jersey 08544-1009, and at Texas A&M University, College Station, Texas 77843-4242, USA.

e-mail: mscully@princeton.edu

1. Bajcsy, M., Zibrov, A. S. & Lukin, M. D. *Nature* **426**, 638–641 (2003).
2. Kocharovskaya, O. *Phys. Rep.* **219**, 175–190 (1992).
3. Arimondo, E. *Prog. Optics* **35**, 257–354 (1996).
4. Harris, S. *Phys. Today* **50**, 36–42 (1997).
5. Scully, M. O. & Zubairy, M. S. *Quantum Optics* (Cambridge Univ. Press, 1997).
6. Hau, L. V., Harris, S. E., Dutton, Z. & Behroozi, C. H. *Nature* **397**, 594–598 (1999).
7. Fitzgerald, R. *Phys. Today* **52**, 17–18 (1999).
8. Boller, K.-J., Imamoglu, A. & Harris, S. E. *Phys. Rev. Lett.* **66**, 2593–2596 (1991).
9. Liu, C., Dutton, Z., Behroozi, C. H. & Hau, L. V. *Nature* **409**, 490–493 (2001).
10. Phillips, D., Fleischhauer, A., Mair, A., Walsworth, R. L. & Lukin, M. D. *Phys. Rev. Lett.* **86**, 783–786 (2001).
11. Kocharovskaya, O., Rostovtsev, Y. & Scully, M. O. *Phys. Rev. Lett.* **86**, 628–631 (2001).
12. André, A. & Lukin, M. D. *Phys. Rev. Lett.* **89**, 143602 (2002).

Chemistry

Cellulose stacks up

Mike Jarvis

The long chains of cellulose pack laterally into microfibrils of two crystalline forms. Comparison of the structures of these two forms reveals unexpected patterns of bonding that tie the chains together.

There is more cellulose on Earth than any other organic substance. It is the main constituent of plant cell walls — and so of the paper on which the printed version of this article appears. For nearly a century the details of its structure have been elusive. But two crystallographic structures, published in *Journal of the American Chemical Society*^{1,2}, now show how cellulose fibres are put together.

There is nothing complicated about a single cellulose molecule; it is an unbranched chain of glucose units linked head to tail. The unsolved question has been how these cellulose chains pack side by side to form microfibrils, linear crystals that in most plants are some 3 nm thick, but which reach widths of more than 20 nm in certain algae and in tunicates or sea squirts, the only animals that make cellulose.

These very large microfibrils are preferred for crystallography, but homogeneity matters as much as size. Native cellulose from almost every source is a mixture of two crystalline forms, I_α and I_β , in statistically variable proportions³. The two forms can occur alongside each other within one of the large algal microfibrils⁴, whereas cellulose from

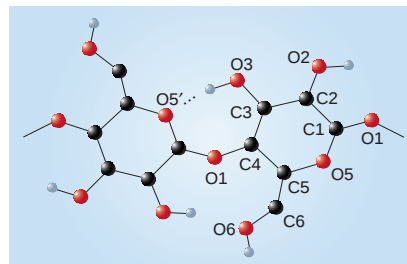


Figure 1 Numbering system for carbon and oxygen atoms in two consecutive glucosyl units of cellulose. The O3–H...O5' hydrogen bond shown is present in all crystalline forms of cellulose, but the pattern of hydrogen bonding from O2 and O6 varies. Hydrogen atoms are shown in grey.

higher plants also contains large numbers of ordered and disordered crystal-surface chains⁵. These problems were bad news for those trying to determine the crystal structure, for they have had to wait until organisms were found that made almost pure cellulose I_α or cellulose I_β .

Nishiyama *et al.*¹ have now published the crystal structure of the I_α form isolated from the freshwater alga *Glaucozystis*. It can be

compared with the structure of cellulose I_β from the tunicate *Halocynthia*, published last year by three of the same authors². Both structures harbour some surprises, and are instructive about the nature of hydrogen bonding.

Many features are common to both crystalline forms. Each cellulose chain approximates to a flat ribbon, with alternate glucose units facing in opposite directions. They are locked in this position by a hydrogen bond between a hydroxyl group (O3–H; Fig. 1) of one glucose unit and the ring oxygen (O5') of the next. All the cellulose chains lie parallel, hydrogen-bonded edge to edge. The sheets of chains so formed are stacked on top of one another with a stagger, along the microfibril, that differs between the I_α and I_β forms.

Most of these features were already known or guessed, but there was a mystery about the forces that held sheets of chains together in a stack. Hydrophobic bonding had been suggested because, with the polar hydroxyl groups ranged along the edges of each ribbon-like chain, its upper and lower faces are relatively nonpolar. But Nishiyama *et al.*¹ point out that the chains are also correctly configured to provide an ordered multiplicity of weak C–H...O hydrogen bonds from one sheet to the next (Fig. 2, overleaf). C–H...O hydrogen bonding has been reported in other biopolymers⁶, but it is remarkable to find it playing such a prominent part in the cohesion of an everyday material such as cellulose.

The I_α and I_β forms were first distinguished by studies using nuclear magnetic resonance⁷. This work showed that each form contained two distinct kinds of glucose unit, more distinct in I_β than in I_α . This is confirmed in the crystal structures^{1,2} and relates elegantly to their lattice symmetry (Fig. 2). In cellulose I_α , alternating glucose units in each chain differ slightly in conformation but all chains are the same. Cellulose I_β has chains of two kinds arranged in alternating sheets. These are termed 'origin' and 'centre' chains from their position in the unit cell. Within each chain, all glucose units are identical, although they face in alternate directions (Fig. 1).

An unexpected feature of both cellulose I_α and cellulose I_β is disorder in the lateral hydrogen bonding that links adjacent chains within a sheet, revealed by analysis of neutron diffraction data with and without substitution of deuterium for the hydroxyl protons^{1,2}. Each crystalline form has two alternative hydrogen-bond networks that differ only in the positions of the hydroxyl protons on O2 and O6. Their occupancy is not in an exact ratio.

In cellulose I_α , adjacent chains are linked by a zig-zag, repeating O–H...O–H... motif¹. In one of the two networks, these strangled hydrogen bonds run continuously along the microfibril towards one end, whereas in the other network they are discontinuous and

run towards the opposite end. The centre chains of cellulose I_{β} have similar hydrogen-bond strings, continuous in each direction². In the origin chains, the hydrogen bonding is either straight across from chain to chain, or solely intramolecular so that adjacent chains are not linked.

The long strings of hydrogen bonds are structurally reminiscent of linear hydrogen-bonding systems seen in protein-bound water ('proton wires')⁷. Cellulose is much more rigid than bound water, though⁵. A question that Nishiyama *et al.*^{1,2} could not answer was whether the alternative hydrogen-bond networks can interconvert cooperatively in time, as required for proton conduction in water⁷, or whether they are permanent but distributed in space within or between microfibrils. This question is fundamental for cellulose I_{β} , where the sheets of origin chains lack cohesion if hydrogen bonding in the centre chains is polarized in one of the two possible directions.

Cellulose is synthesized at the cell surface by membrane-spanning synthase complexes. These take the form of regular arrays of particles in most algae⁸ and hexameric rosettes in higher plants, with each rosette containing cellulose synthase enzymes of three distinct kinds⁹. Is it the geometry of these arrays and rosettes, or self-assembly, that determines the structure of cellulose microfibrils? Both, it seems. The lateral dimensions of microfibrils vary with the size of the arrays^{8,10}, and in each microfibril the chains are parallel (heads all pointing the same way) because that is how they emerge from the synthetic complex.

How the terminal complexes fit together is influenced by the sequence of the cellulose synthases^{9,10}. If the internal structure of each microfibril were templated by its terminal complex, we should expect aberrant crystalline cellulose structures to result from mutations disrupting the complex. So far no such aberrations have been described, suggesting that the cellulose I_{α} and I_{β} lattices self-assemble. But there is a twist to this tale. In the model plant *Arabidopsis*, a single amino-acid substitution in one of the cellulose synthases interferes with rosette assembly and is sufficient to disrupt ordered orientation of the microfibrils¹¹. The resulting plants are dwarfed, apparently through disorientation of microfibrils that would normally be wrapped around the cells of plant stems to constrain them into elongating rather than expanding equally in all directions. This is not the only factor influencing the direction of growth, but normal synthesis of cellulose seems to be needed for the normal development of plant form.

The new structures do not close the debate on the nature of cellulose. Is their unexpected complexity random, the consequence of imprecise self-assembly? Or is it the result of spatial patterning by the geometry of the

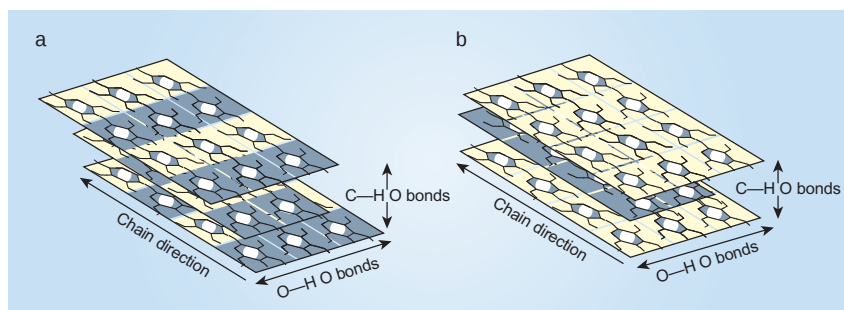


Figure 2 Symmetry and directions of hydrogen bonding in cellulose^{1,2}. **a**, Cellulose I_{α} , in which all chains are crystallographically identical but alternating glucose units in each chain, shaded grey and yellow, differ slightly in conformation. **b**, Cellulose I_{β} , in which chains of two distinct kinds are arranged in alternating sheets. Chains passing through the origin and centre of the unit cell are shaded respectively yellow and grey.

terminal complexes, so that understanding structural detail and biosynthesis will have to go hand in hand? Could the hydrogen-bond networks be switchable? If they are, cellulose is a smarter material than anyone thought. ■

Mike Jarvis is in the Department of Chemistry, University of Glasgow, Glasgow G12 8QQ, UK. e-mail: mikej@chem.gla.ac.uk

1. Nishiyama, Y., Sugiyama, J., Chanzy, H. & Langan, P. J. *Am. Chem. Soc.* **125**, 14300–14306 (2003).
2. Nishiyama, Y., Chanzy, H. & Langan, P. J. *Am. Chem. Soc.* **124**, 9074–9082 (2002).

3. Atalla, R. H. & VanderHart, D. L. *Science* **223**, 283–285 (1984).
4. Sugiyama, J., Vuong, R. & Chanzy, H. *Macromolecules* **24**, 4168–4175 (1991).
5. Viñtor, R. J., Newman, R. H., Ha, M. A., Apperley, D. C. & Jarvis, M. C. *Plant J.* **30**, 721–731 (2002).
6. Babu, M. M., Singh, S. K. & Balaram, P. J. *Mol. Biol.* **322**, 871–880 (2002).
7. Cui, Q. & Karplus, M. J. *Phys. Chem. B* **107**, 1071–1078 (2003).
8. Koyama, M., Sugiyama, J. & Itoh, T. *Cellulose* **4**, 147–160 (1997).
9. Taylor, N. G. *et al.* *Proc. Natl Acad. Sci. USA* **100**, 1450–1455 (2003).
10. Roberts, A. W., Roberts, E. M. & Delmer, D. P. *Eukaryot. Cell* **1**, 847–855 (2002).
11. Sugimoto, K., Williamson, R. E. & Wasteneys, G. O. *Protoplasma* **215**, 172–183 (2001).

Geomorphology

Nature, nurture and landscape

Peter Molnar

Those studying erosion in mountain regions wrestle with factors such as what builds mountains, and how climate affects erosive forces. Yet perhaps a physically based theory is what is most needed.

The endless debate over the relative importance of nature and nurture in child development has its equivalent in geomorphology. In this case, the argument is about the roles of tectonics and climate in mountain erosion. Tectonics (nature) sets the initial conditions by raising Earth's surface and, where active, renewing topography. Climate (nurture) shapes the surface into its various forms through its effect on glaciers and rivers. Three papers in this issue^{1–3} and another in *Geology*⁴ take the argument forward by isolating and evaluating the importance of certain climatic and tectonic factors in erosion.

As well as offering individual insights, these papers show how lack of resolution in the debate is impelling research in geomorphology away from observation and towards theory. Moreover, in different ways, they remind us that understanding the physical processes that govern the most pervasive of erosive forces, river erosion, requires a geological approach to provide data way back in time. Although there is a long tradition

of measuring sediment transport by rivers, only infrequently will that record include the full range of floods that have shaped the landscape. Flooding obeys a power-law distribution⁵, and the largest floods have the largest effect on landscape. But only in rare regions are we likely to have witnessed them — hence the need for an approach that will produce a record that includes several such extreme events.

To quantify erosion rates, all four groups^{1–4} apply thermochronometric methods to rock now exposed at the surface. Temperature increases with depth in the Earth and, at high temperatures, noble gases emitted in radioactive decay diffuse away; defects (tracks) in crystals, produced by the charged particles expelled in nuclear fission, anneal. By measuring the concentrations of such gases or tracks, one can date when the sample cooled below a temperature at which diffusion or annealing is slow. With an assumed temperature gradient in the Earth, the upshot is an estimate of the average exhumation rate, which in these cases equals the rate

that material above the sampled rock has been eroded. Such estimates apply to periods as short as 500,000 years to as long as several million years. But in all cases they span several glacial and interglacial cycles, and so smooth the effects of large climatic changes.

Reiners *et al.*¹ present the simplest result: erosion rates averaged over the past few million years in the North Cascade Mountains of Washington state correlate with the present-day distribution of rainfall. Precipitation and erosion rates vary by an order of magnitude across the range, with rapid erosion having occurred where rain falls most today. As virtually all of the rock exposed in the mountains must have lain below sea level a few million years ago, they deduce that rock moved up relative to sea level despite the absence of any tectonic process. Because of isostasy (Archimedes' principle applied to the Earth's crust immersed in its more dense mantle)^{1,6}, removal of a mass of rock from the Earth's surface will be compensated by the rise above sea level of approximately 80–85% of that mass.

By contrast, Burbank *et al.*² deny that precipitation has a major role in erosion in their study area, the Himalaya. They measured rainfall along a profile across a segment of the Himalaya, and they compare that rainfall with estimates of erosion rates along a valley floor from the Lesser Himalaya into the Greater Himalaya. A narrow zone separates the Lesser Himalaya (where wide rivers, bounded by sediment-filled terraces, flow through relatively gentle, deeply weathered hillslopes) from the snow-capped Greater Himalaya (where slopes are steep, valleys deep, and weathering and terraces sparse).

Burbank *et al.* measured much more rapid erosion in the Greater than the Lesser Himalaya. More importantly, they detected no measurable difference in erosion rates across the Greater Himalaya despite a five-fold decrease in precipitation there. They

infer that precipitation does not exert a first-order control on erosion. Noting that all of the rapidly eroding terrain moves rapidly upward with respect to the lower, gentler region to the south, and leaving open the question of what physical processes cause erosion, Burbank *et al.* suggest that tectonically forced upward movement is the most important factor affecting erosion across a region of such different rainfall.

Using a different thermochronometer, spanning several million years, Wobus *et al.*⁴ deduce average erosion rates in a neighbouring part of Nepal. Like Burbank *et al.*, these authors report a large difference between high erosion rates in the Greater Himalaya and low rates in the Lesser Himalaya. Because their thermochronometer applies to a much longer period of time, their results require a larger difference in the amount of rock eroded. Only a few kilometres of rock have been removed from the Lesser Himalaya since the Himalaya began to form some 40–50 million years ago, but in the Greater Himalaya roughly 10 km have been removed since about 10 million years ago. There is no obvious fault or shear zone between them, but Wobus *et al.* logically deduce that the rock in the Greater Himalaya has moved upwards relative to that in the Lesser Himalaya. Whereas Burbank *et al.* attribute the constant erosion rates across the Greater Himalaya to rapid rates of vertical movement of the rock, Wobus *et al.* suggest the opposite: that the rapid rise of Greater Himalayan rock results from its rapid erosion and isostatic compensation of the mass removed.

Common to these three papers^{1,2,4} is the sensible assumption that the landscapes studied have reached some form of equilibrium, so that the basic character of the landscape and the rates at which geomorphological processes have shaped it have not changed over the time spanned by the measured erosion.

For their part, Dadson *et al.*³ examined



Figure 1 Peak erosion — one of the areas of Taiwan studied by Dadson *et al.*³. They conclude that these peaks have been stripped of material by earthquake-triggered landslides. (Photo courtesy of J. C. Lin, National Taiwan Univ.)



100 YEARS AGO

For some years a very interesting series of experiments in connection with the biological method of sewage treatment has been carried on by Dr. Dunbar, director of the Hygienisches Institut at Hamburg, and by his colleagues. Special attention has been directed to the elucidation of the sequence of changes which underlies the purification process in contact beds and percolating filters... Great importance is attached by the Hamburg workers to the role played by the process of so-called "absorption" which takes place when the liquid is in contact with the purifying medium. It has been found that sterile clinkers have the power of withdrawing from solution not only colouring matters, but also the highly complex nitrogenous bodies found in sewage... An interesting example of absorption is seen in the case of the percolating filter adopted by Dr. Dunbar. This filter is provided with a layer of fine material on the surface about six inches deep. According to Dr. Dunbar, 50 per cent. of the purification, apart from nitrification, takes place in this six inches.

From *Nature* 10 December 1903.

50 YEARS AGO

In case there is any lingering doubt that the Piltdown finds are in part fraudulent, we think that one other fact now brought to light should be published immediately. Suspecting that some of the so-called implements reported from the site might have been 'doctored', we asked Mr. E. T. Hall, of the Clarendon Laboratory, Oxford, to test the composition of their surface stains by means of his X-ray spectrographic method of analysis. He has reported to us that the stains on these flints are entirely ferruginous, with one notable exception. The triangular flint (Reg. No. E.606) recovered *in situ* from the layer immediately overlying the skull horizon is chromate stained. When this stain is removed in acid the flint appears greyish-white. It is indistinguishable from a mechanically broken piece of flint such as one might encounter on the surface of any ploughed field in 'Chalk-land'. Whereas a bone might have been dipped in a solution of potassium dichromate with the sole purpose of trying to harden it, a flint would only have been treated in that way by a forger requiring it to be of a certain colour.

From *Nature* 12 December 1953.

erosion of Taiwan at three different timescales. They find both commonality and marked differences in their results. On the oldest rock, where topography is steep, the high erosion rates averaged over the past few million years differ little from those inferred from modern (the past 30 years or less) rates, obtained from ratios of rates of sediment transport by rivers to the areas of their watersheds. On the flanks of the mountainous terrain, where folding and faulting at depth build new topography in weaker rock, present-day erosion rates are highest, despite the low stream gradients and gentle hillslopes, characteristic of slow erosion elsewhere. Moreover, for some of these areas, the average erosion rate obtained from 30 years or so of data exceeds the rates of river incision of narrow valleys over the past 10,000 years. Thus, the modern average erosion rates in these regions cannot have applied to a geological period as short as 10,000 years.

Dadson *et al.* searched for spatial correlations of erosion with factors such as precipitation rate, river discharge, stream gradient and stream power (see below). They conclude that only two correlate well: recent seismicity (Fig. 1), and precipitation associated with typhoons. Earthquakes and large storms are notorious triggers of landslides, which abruptly carry debris to rivers; so these correlations are less surprising than the failure of the other factors. Moreover, the historical record of seismicity on Taiwan goes back only about 100 years, and thus is surely too short to yield a representative image of the distribution of landslide triggers. Perhaps it is no wonder that modern average erosion rates and those averaged over geological times differ in the low parts of Taiwan.

To find a unifying concept for erosion rates, many have turned to stream power. In a precocious effort to understand how running water erodes very weak rock, Howard and Kerby⁷ suggested that erosion rates should vary with the stress a flowing stream exerts on its bed, which can be expressed in terms of the discharge and the river gradient, or stream power per unit width. Accordingly, much present work addresses the role of stream power in erosion⁸. Surely the greater the discharge, the faster material can be removed, and abundant evidence shows a correlation of present-day average erosion rates with the steepness of terrain^{9,10}.

As far as the new work is concerned, Reiners *et al.*¹ and Dadson *et al.*³ find no correlation between their measured erosion rates and stream power. Burbank *et al.*² show that the large variation in rainfall across the Himalaya is compensated to some extent by steeper gradients where rainfall is low, but not sufficiently for them to embrace stream power as the key to account for the poor correlation of erosion rate with rainfall. At the other extreme, Wobus *et al.*⁴ exploit the large contrast in stream power between

the Greater and Lesser Himalaya as support for the more rapid erosion in one than the other.

Water is Earth's universal solvent, and without it erosion would slow. Yet, as a solution to what makes erosion fast or slow, water's role remains controversial. The differences among these papers call attention to the inadequacy of current theory, without which one gropes for a way to plot data. ■

Peter Molnar is in the Department of Geological Sciences, and the Cooperative Institute for Research in Environmental Science (CIRES), University of Colorado at Boulder, Boulder, Colorado 80309-0399, USA.

e-mail: molnar@cires.colorado.edu

1. Reiners, P. W., Ehlers, T. A., Mitchell, S. G. & Montgomery, D. R. *Nature* **426**, 645–647 (2003).
2. Burbank, D. W. *et al.* *Nature* **426**, 652–655 (2003).
3. Dadson, S. J. *et al.* *Nature* **426**, 648–651 (2003).
4. Wobus, C. W., Hodges, K. V. & Whipple, K. X. *Geology* **31**, 861–864 (2003).
5. Turcotte, D. L. & Greene, L. *Stochastic Hydrol. Hydraul.* **7**, 33–40 (1993).
6. Holmes, A. *Principles of Physical Geology* 189–190 (Ronald Press, New York, 1944).
7. Howard, A. D. & Kerby, G. *Geol. Soc. Am. Bull.* **94**, 739–752 (1983).
8. Whipple, K. X. & Tucker, G. E. *J. Geophys. Res.* **104**, 17661–17674 (1999).
9. Ahnert, F. *Am. J. Sci.* **268**, 243–268 (1970).
10. Montgomery, D. R. & Brandon, M. T. *Earth Planet. Sci. Lett.* **201**, 481–489 (2002).

Signal transduction

Molecular monogamy

Drew Endy and Michael B. Yaffe

The interactions between cellular proteins must be highly specific, or cells will stop functioning. Some clever protein-manipulation experiments have revealed how this specificity has evolved in yeast.

If an integrated system is to function correctly, its components must be wired together accurately. This requirement presents a particular challenge for living cells, because cellular components move about and intermix, and because the 'wires' themselves are dynamic molecular interactions. In many cells, for instance, protein-based signal-transduction systems are assembled through protein–protein interactions. These interactions are often specified by structurally defined 'domains' in one protein that bind to complementary short, linear amino-acid sequence motifs in another. But even a relatively simple cell such as baker's yeast normally produces more than 4,500 different proteins¹, and frequently several of these proteins contain similar domains or motifs. How, then, do cells wire proteins together with high specificity? On page 676 of this issue, Zarrinpar and colleagues² describe one way in which yeast ensures a monogamous protein partnership — by eliminating nonspecific interactions through evolution.

For some types of interaction domains, such as the so-called SH2 and FHA domains, two mechanisms contribute to binding specificity. First, binding occurs only when a particular tyrosine, serine or threonine amino acid in the partner motif has been enzymatically tagged with a phosphate group. The phosphate contributes a large fraction of the total energy required for motif–domain binding³. Second, additional amino acids flanking the phosphorylated residue fine-tune the interaction, discriminating between specific and nonspecific partners. But what about other modular domains that recognize more promiscuous

motifs, with considerably lower affinity? For instance, how is specificity obtained for SH3 domains, which recognize the core sequence motif proline–X–X–proline (where X is any amino acid)?

This is where the findings of Zarrinpar *et al.*² come in. These authors chose to study the interaction between two proteins, Sho1 and Pbs2, from the high-osmolarity glycerol (HOG) signalling pathway in baker's yeast (*Saccharomyces cerevisiae*)⁴. Sho1 is a sensor protein that sits in the membrane of yeast cells and detects changes in external osmolarity. Pbs2 is a signalling protein that coordinates the cellular response to Sho1 activation; the resulting changes in glycerol production help to balance the intracellular and external osmotic pressures. Sho1 and Pbs2 connect through a domain–motif interaction: Sho1 contains an SH3 domain and Pbs2 contains an SH3-binding motif. The SH3 domain in Sho1 is one member of a broader family — there are 27 known SH3 domains in *S. cerevisiae* proteins⁵. But HOG-pathway signalling depends on the specific interaction between Sho1 and Pbs2; cross-reaction with any component from the other 26 SH3-domain–motif pairs could gum up the inner workings of the cell.

Zarrinpar *et al.*² start by showing that this doesn't happen (Fig. 1a). They constructed 38 artificial Sho1 proteins by replacing the native Sho1 SH3 domain with one of the 26 other yeast SH3 domains, or with one of 12 such domains taken from multicellular organisms. None of the Sho1 proteins created from the 26 alternative yeast SH3 domains could reconstitute HOG-pathway function *in vivo*. Curiously, however, six of

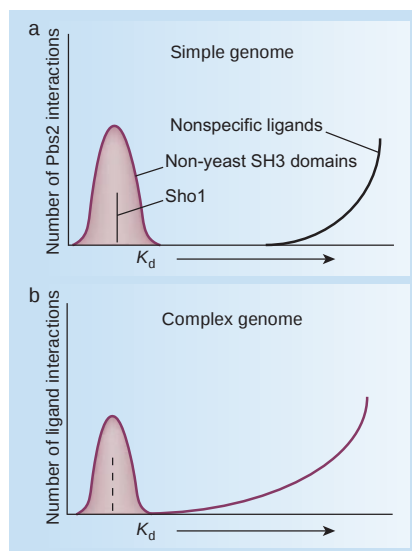


Figure 1 How organisms achieve specificity of protein–protein interactions. A general feature is the presence of structural ‘domains’ in one protein and complementary ‘motifs’ in their binding partners (ligands). **a**, In baker’s yeast, which has a relatively simple genome encoding relatively few proteins, a single motif-containing protein (here, Pbs2) binds to just one SH3-domain-containing partner (Sho1) with a reasonably low dissociation constant (K_d ; that is, with high affinity). Other yeast SH3-domain proteins bind to Pbs2 with lower affinity (black curve). Zarrinpar *et al.*² find that this specificity results from evolutionary negative selection against nonspecific interactions. SH3 domains from other organisms are not subject to negative selection in yeast, and so bind promiscuously to Pbs2 with dissociation constants similar to that of Sho1 (purple curve). **b**, In organisms with more complex genomes, which encode many SH3 domains and many ligands that bind these domains, additional mechanisms may work to restrict a large number of potential interactions (purple curve) to a single domain–ligand pair (dashed line).

the 12 non-yeast SH3 domains functioned well enough to allow cell growth under high salt conditions (where the HOG pathway is important). Why should this be? Zarrinpar *et al.* suggest that, through natural selection, the amino acids within and around the proline–X–X–proline motif on Pbs2 have evolved to be recognized only by the SH3 domain of Sho1, and not by any other yeast SH3 domain. No such negative selection would have occurred against the non-yeast SH3 domains. In general terms, then, an evolving system composed of intermixing parts could use negative selection to eliminate spurious interactions.

To test this model, the authors changed the Pbs2 SH3-binding motif so as to increase or decrease the strength of the Sho1–Pbs2 interaction. All changes reduced the specificity of interaction for the Sho1 SH3

domain, suggesting that the Pbs2 motif was already ‘optimized’ for the combination of binding strength and SH3 specificity. As a second test, Zarrinpar *et al.* used a competitive growth assay to compare yeast containing wild-type Pbs2 with yeast containing either a mutant Pbs2 that does not interact with Sho1, or a promiscuous Pbs2 mutant that interacts with both Sho1 and most other yeast SH3 domains. Both the wild-type and the promiscuous strains outgrew the ‘non-interacting’ strain under high salt conditions. But the wild-type and non-interacting strains outgrew the promiscuous strain under conditions that do not require the HOG pathway. The success of the non-interacting strain under these conditions supports the general model that a cell is better off with components that don’t interact than with those that bind to one another indiscriminately. Promiscuous proteins in a particular cell type are selected against in order to maintain a ‘self-consistent’ protein–interaction network that is free of detrimental interactions.

What happens in more complex animals, in which there are greater numbers of protein–protein interactions (Fig. 1b)? Is negative selection alone sufficiently powerful to maintain interaction specificity? The data here are incomplete. Analysis of some mammalian SH3 domains by techniques such as phage display and bioinformatics suggests more promiscuous protein binding, and a lack of strong, specific motif selection⁵. But some of this apparent promiscuity is clearly overcome by temporal and spatial segregation; not all components are present at the same time and place. So negative selection such as that described by Zarrinpar *et al.* need only operate within the confines

of a particular subcellular compartment or cell type in higher animals.

In addition, these organisms may have developed further mechanisms for maintaining interaction specificity. Genomic analyses of multicellular organisms suggest the evolution of more complex multidomain protein architectures^{6,7}, in which SH3 domains are mixed-and-matched with other modular domains. This might allow a series of weak and otherwise promiscuous individual interactions with any potential target protein to occur simultaneously, with the sum of these interactions providing much higher specificity than that possible with any single domain acting alone.

Zarrinpar and colleagues’ work² reveals an elegant example of how biology has solved the problem of wiring dynamic systems at the molecular scale. A complete understanding of such systems, and of the mechanisms that underlie their proper function and maintenance, will greatly aid our analysis of — and interaction with — the living world. ■

Drew Endy and Michael B. Yaffe are in the Department of Biology and the Division of Biological Engineering, and Michael B. Yaffe is at the Center for Cancer Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139-4307, USA. e-mail: myaffe@mit.edu

1. Ghaemmaghami, S. *et al.* *Nature* **425**, 737–741 (2003).
2. Zarrinpar, A., Park, S.-H. & Lim, W. A. *Nature* **426**, 676–680 (2003).
3. Bradshaw, J. M., Mitaxov, V. & Waksman, G. J. *Mol. Biol.* **293**, 971–985 (1999).
4. O’Rourke, S. M., Herskowitz, I. & O’Shea, E. K. *Trends Genet.* **18**, 405–412 (2002).
5. Brannetti, B., Via, A., Cestra, G., Cesareni, G. & Helmer-Citterich, M. J. *Mol. Biol.* **298**, 313–328 (2000).
6. International Human Genome Sequencing Consortium *Nature* **409**, 860–921 (2001).
7. Venter, J. C. *et al.* *Science* **291**, 1304–1351 (2001).

Chirality

Organic films with a twist

Michael D. Ward

Left- and right-handed helical molecules form mirror-image chiral crystals on a copper substrate. It seems that the substrate and the molecules work in concert to determine the handedness of the crystal domains.

Chirality is central to the building blocks of life, and to commercial chemical enterprises. Most amino acids, sugars and pharmaceuticals contain chiral carbon centres — a carbon atom bonded to four different substituents in a tetrahedral geometry. Such chiral molecules exist in two mirror-image forms, like left and right human hands, that are called enantiomers. Our understanding of this peculiar property can be traced back to Louis Pasteur, who discovered that ‘racemic acid’ (a crystalline deposit formed on wine casks during fermentation) consisted of equal amounts of

left- and right-handed crystals of sodium ammonium tartrate, which were easily distinguished as mirror images under an optical microscope¹. As they report in *Angewandte Chemie International Edition*, Fasel *et al.*² have exploited the atomic-level imaging capabilities of the scanning tunnelling microscope (STM) to observe chirality directly at the molecular level, in enantiomorphic two-dimensional crystals of chiral molecules on a copper surface.

In three dimensions, chiral molecules can form either racemic (heterochiral) crystals, which contain equal numbers of left- and

right-handed molecules in the same crystal, or conglomerates of separate left- or right-handed (homochiral) crystals of the pure enantiomers, like Pasteur's tartrate salts. Racemic crystals, however, tend to greatly outnumber conglomerates³. Despite the passage of 150 years since Pasteur's studies, the factors responsible for the transmission of chiral information between molecules during crystal formation are not fully understood. Nevertheless, theory has suggested that discrimination between hetero- and homochiral ordering becomes more likely when short-range repulsive forces (such as those experienced by molecules in close-packed organic crystals) are significant⁴. Under these conditions, the selectivity for racemates or conglomerates would be governed by molecular shape and symmetry.

How chiral structures form on surfaces, and how chiral information propagates from single molecules to supramolecular ensembles, are important questions in several technological applications, including crystal growth^{5,6}, the fabrication of liquid-crystal displays⁷ and enantioselective catalysis^{8,9}. Like many organic molecules, chiral molecules can adsorb on crystalline substrates, such as graphite or copper, to form two-dimensional crystals that are only a single molecule thick. If the substrate is electrically conductive, the adsorbed molecules can be viewed with an STM. To create an image, an ultrasharp electrified tip is scanned over the crystals and the tunnelling current measured between the tip and crystals on the conductive substrate.

The extraordinary spatial resolution and the sensitivity of the tunnelling current to the different chemical groups in a molecule means that the handedness of chiral structures can be assigned¹⁰, as can their epitaxial alignment on the substrate¹¹. Scanning tunnelling microscopy has revealed that, unlike their three-dimensional counterparts, racemic compounds often form conglomerates of two-dimensional enantiomorphic crystals when confined to a substrate surface^{12,13}, suggesting that confinement in two dimensions aids chiral discrimination¹⁴. These enantiomorphic two-dimensional crystals are discernible as mirror images¹⁵, not unlike Pasteur's tartrates.

Ernst and co-workers previously found¹⁶ that racemic mixtures of '[7]H' — which consists of seven benzene rings fused edge-to-edge to form a chiral helical coil — formed separate enantiomorphous domains aligned through epitaxy on a Cu(111) substrate (the Cu(111) surface is a flat plane of copper atoms in a hexagonal arrangement). Now Fasel *et al.*² have prepared films of the pure enantiomers, designated (M)-[7]H and (P)-[7]H, on the Cu(111) surface. STM images revealed spheres corresponding to individual molecules of [7]H organized into two-dimensional crystalline monolayers. When

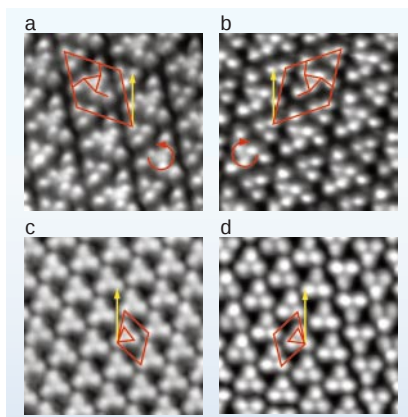


Figure 1 Left- and right-handed forms of [7]H. a, b, The '6&3' structure; c, d, the '3-structure'. Unit cells and the arrangement of molecules within them are outlined in red; the yellow arrow indicates the same surface direction in each figure.

95% of the surface was covered with [7]H, the molecules organized as pairs of triangular clusters, one with three molecules and the other with six (a '6&3' structure; Fig. 1a, b). Surprisingly, increasing the coverage just slightly, to 100%, produced a different film structure, consisting of three-molecule clusters (a '3-structure'; Fig. 1c, d). The chirality of these films is apparent from the STM images — the enantiomeric (M)-[7]H and (P)-[7]H lattices are mirror images aligned along opposite directions on the substrate surface. These features signify epitaxial ordering and, more importantly, they reveal that the chirality of the [7]H molecules is transmitted to the molecule–substrate interface.

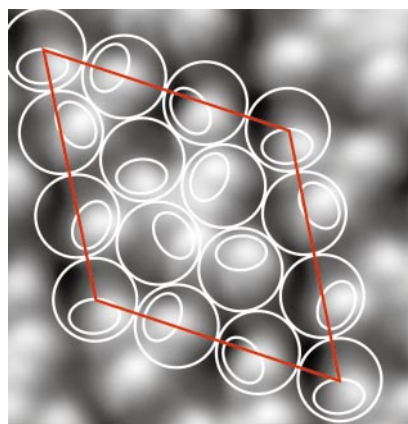


Figure 2 Chirality established. Fasel *et al.*² have overlaid their STM image of (M)-[7]H molecules with a model of the structure of these 6&3 molecule clusters. Each circle represents a molecule, and the bright oval inside each sphere is the topmost part of the molecule. Tracing a path between them, each molecule is rotated through 60° with respect to the previous one, indicating that chirality is transmitted between adjacent molecules.

Particularly striking is the tunnelling contrast observed in the 6&3 clusters. The six-membered clusters resemble a 'pinwheel', with the (M)-[7]H pointing anti-clockwise and (P)-[7]H pinwheel clockwise, a direct consequence of the chirality of the helical molecules. Zooming in on the images reveals that each molecular sphere has a circular bright spot tipped to one side (Fig. 2). The authors ascribe this feature, quite reasonably, to tunnelling at the uppermost ring of the [7]H molecule, which is closest to the STM tip. Remarkably, tracing an anti-clockwise circuit on the periphery of the (M)-[7]H six-membered cluster, molecule by molecule, reveals that this bright spot rotates in successive steps by 60° about the surface normal. This produces a total of six different in-plane molecular orientations. A three-molecule circuit around the three-membered cluster of a 6&3 structure or a 3-structure cluster also reveals successive 120° orientations of the bright spot.

The observation that successive 60° (or 120°) orientations are observed during these circuits, instead of random 60° orientations, can only mean that chirality is transmitted between adjacent [7]H molecules. This [7]H is not decorated with strong dipoles or hydrogen-bonding groups that could direct molecular organization. Instead, it would seem that the helical shape of [7]H prescribes a gear-like rotation of adjacent molecules in these close-packed two-dimensional crystals. Fasel *et al.*² have thus provided a molecular-level insight into how chirality evolves in crystals, which will undoubtedly enhance our understanding of chirality in supramolecular ensembles.

Michael D. Ward is in the Department of Chemical Engineering and Materials Science, University of Minnesota, 421 Washington Avenue, Minneapolis, Minnesota 55455-0132, USA.

e-mail: wardx004@umn.edu

- Pasteur, L. *Ann. Chim. Phys.* **24**, 442–459 (1848).
- Fasel, R., Parschau, M. & Ernst, K.-H. *Angew. Chem. Int. Edn* **42**, 5178–5181 (2003).
- Jacques, J., Collet, A. & Wilen, S. H. *Enantiomers, Racemates and Resolutions* (Wiley, New York, 1981).
- Schipper, P. E. & Harrowell, P. R. *J. Am. Chem. Soc.* **105**, 723–730 (1983).
- Weissbuch, I., Addadi, L., Lahav, M. & Leiserowitz, L. *Science* **253**, 637–645 (1991).
- Orme, C. A. *et al. Nature* **411**, 775–779 (2001).
- Oda, M. *et al. Adv. Mater.* **12**, 362–365 (2002).
- Switzer, J. A., Kothari, H. M., Poizat, P., Nakanishi, S. & Bohannan, E. W. *Nature* **425**, 490–493 (2003).
- Humboldt, V., Haq, S., Muryn, C., Hofer, W. A. & Raval, R. *J. Am. Chem. Soc.* **124**, 503–510 (2002).
- Stevens, F., Dyer, D. J., Muller, U. & Walba, D. M. *Langmuir* **12**, 5625–5629 (1996).
- Hooks, D. E., Fritz, T. & Ward, M. D. *Adv. Mater.* **13**, 227–241 (2001).
- DeFeyer, S. *et al. Angew. Chem. Int. Edn* **40**, 3217–3220 (2001).
- DeFeyer, S. *et al. Langmuir* **16**, 9887–9894 (2000).
- Stewart, M. V. & Arnett, E. M. in *Topics in Stereochemistry* (eds Allinger, N. L., Eliel, E. L. & Wilen, S. H.) 195–262 (Wiley, New York, 1982).
- Xu, Q.-M. *et al. Angew. Chem. Int. Edn* **41**, 3408–3411 (2002).
- Ernst, K.-H., Kuster, Y., Fasel, R., Muller, M. & Ellerbeck, U. *Chirality* **13**, 675–678 (2001).

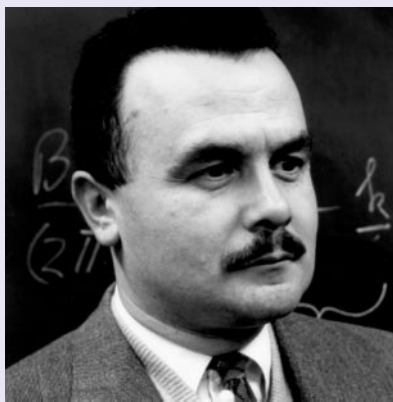
Obituary

Bertram N. Brockhouse (1918–2003)

Advances in science often start with abstract and difficult ideas — Mendel's gene or Dalton's atom come to mind. Only much later, through defining experiments, do they become such concrete objects that one wonders what all the fuss was about. Genes have become sequences of nucleic acids, and atoms are now real things that can be manipulated one at a time with scanning probe microscopes. So it is, too, in condensed-matter physics: the phonon and the magnon were postulated as elementary quantum excitations of matter but acquired a real existence only through a brilliant set of experiments with thermal neutrons in the late 1950s. Those experiments were performed by Bertram Brockhouse and marked the beginning of a new field of spectroscopy, inelastic neutron scattering. Brockhouse — who shared the 1994 Nobel Prize in Physics with Clifford Shull — died on 13 October 2003.

In the 50 years since Brockhouse's invention, neutron spectroscopy has become a standard tool of physics. The neutron, an electrically neutral particle, is able to penetrate deep in the interior of matter; it possesses a magnetic moment and has, at room temperature, a wavelength that nicely matches the distance between atoms in solids and liquids. Dozens of reactors, and more recently accelerators, have been built for neutron spectroscopy: the new Spallation Neutron Source in Oak Ridge, Tennessee, will provide the most intense pulsed neutron beams ever used in research, and, at a total cost of US\$1.4 billion, is one of the largest scientific projects in the world.

Brockhouse was raised on a farm in Alberta but showed an early interest in electronics, running a small radio repair business before joining the Canadian Navy at the start of the Second World War. After physics degrees from the Universities of British Columbia and Toronto, in 1950 Brockhouse joined the Chalk River Nuclear Laboratory in Toronto. Encouraged by Don Hurst there, who had already built a prototype neutron spectrometer, he started building a series of increasingly sophisticated neutron spectrometers to study excitations in solids. Unlike Shull, who worked on the static positions of atoms and spins, Brockhouse focused on their movements. For this, he needed to measure the neutrons' energy change as they scattered from his sample, tying down the incoming and outgoing neutron energies, as well as the momentum transfer ('Q') from the neutron to an excitation in the sample.



Father of neutron spectroscopy

By 1957 his famous triple-axis spectrometer, operating in the 'constant Q' mode, was complete. The first axis held a crystal used to select the incoming neutron energy; the second, the sample under study; and the third axis analysed the scattered neutron energy. The choice of angles of the axes was determined by simple energy- and momentum-conservation rules: set all the angles in such a way as to keep Q (the momentum transferred to the crystal) constant while varying the energy change of the neutron. Taking advantage of the high neutron flux of the new National Research Universal nuclear reactor at Chalk River, then the world's most powerful source of neutrons, Brockhouse wasted no time in applying his new invention to a series of fundamental experiments. In rapid succession, he and his co-workers produced ground-breaking papers on phonons in metals, semiconductors and insulators; on magnons in magnetic materials; and on diffusive molecular and atomic motions in liquids.

The phenomenon of superconductivity offers a good example of the impact that neutron spectroscopy had on physics. About the same time that the detailed phonon spectra of metals were emerging from Chalk River, Bardeen, Cooper and Schrieffer had announced their theory of superconductivity. The star actor in the theory was the phonon; it was the glue that bound the electrons into pairs to form the superconducting state at low temperatures. It is fair to say that it was the exquisite agreement between neutron spectroscopy and superconducting tunnelling experiments on the self-energies of the electrons that convinced the last sceptics that the problem of low-temperature

superconductivity had finally been solved. The same debate is currently raging in the context of high-temperature superconductivity, where neutron spectroscopy is again a central player in determining the nature of the force that binds the superconducting electrons. Magnon and phonon spectra have been measured with neutrons and are being correlated with spectra directly related to the forces between the charge carriers.

In 1962 Brockhouse moved to McMaster University in Hamilton, Ontario — a good choice, as it had an on-campus nuclear reactor with a fairly substantial neutron flux. Brockhouse quickly gathered an enthusiastic group of graduate students and faculty interested in the new spectroscopy. He built a triple-axis spectrometer at the reactor to give his graduate students hands-on training, although for serious research they still had to travel to one of the large national laboratories. At McMaster, Brockhouse's interest shifted gradually away from neutrons towards broader issues. As chair of the Department of Physics he introduced several new programmes, one of which, astronomy, has blossomed over the years so much that the department's name has been changed to Physics and Astronomy. However, the Brockhouse neutron legacy remains. The nuclear reactor is still running, and Bruce Gaulin, holder of the Brockhouse Chair in Physics of Materials, is leading the efforts to build a Canadian beamline at the Spallation Neutron Source in Oak Ridge.

Today, the phonon and the magnon are seen as fundamental elements of matter almost on the level of atoms or molecules. But Brockhouse himself had trouble fully accepting the reality of the objects that he had brought to life. In his Nobel prize lecture, he talked of the "grand atlas of the physical world", in which atoms and their positions have a place. But he was not sure that the phonon and the magnon belonged in it. This scepticism may have been grounded in a general mistrust of quantum phenomena that was still common in the 1940s, when Brockhouse received his graduate education. A more likely cause for his reserve rests deeper — Bertram Brockhouse was a very modest man. But perhaps he was ready to accept the phonon and magnon, after all. Addressing a group of undergraduate students after winning his Nobel prize, he said: "I used to think that my work was not important, but recently I have had to change my mind." Tom Timusk

Tom Timusk is in the Department of Physics and Astronomy, McMaster University, Hamilton, Ontario L8S 4M1, Canada.
e-mail: timusk@mcmaster.ca

Metallurgy

Cost and compromise in alloy engineering

Appl. Phys. Lett. **83**, 4527–4529 (2003)

There's such a thing as too much choice, as materials scientists searching for new alloys know. There is a phenomenal number of ways of combining four or more metallic elements. So an experimental search for the ideal alloy for a particular application is often impractical. Thomas Bligaard and colleagues propose an alternative. They have used first-principles computational methods to predict relevant properties, such as heats of formation and bulk moduli (in essence, compressibility), for more than 64,000 different alloys. Such calculations are now accurate enough to be considered 'virtual experiments'.

But choosing an engineering material is rarely about optimizing a single property. A common compromise is that between performance and cost. Bligaard *et al.* say that such a compromise is similar to the balance sought between utility, cost and risk in economics, and they suggest that it can be found using the notion of Pareto optima: solutions that cannot be improved in one respect without degrading them in another. They show how to find the Pareto-optimal alloys that balance compressibility against cost. Low compressibility implies not only hardness but also often low thermal expansion, which is needed for precision engineering. The authors suggest new mid-priced candidate alloys such as ruthenium silicide.

Philip Ball

Meteorite impacts

Dinosaurs were not roasted

Geology **31**, 1061–1064 (2003)

Wildfires did not ravage the planet after the giant meteorite impact at the end of the Cretaceous period, according to Claire M. Belcher and co-workers.

They have looked at North American sedimentary rocks dating from the Cretaceous–Tertiary (K/T) boundary, 65 million years

ago, searching for signs of charcoal that signify ancient wildfires. Whereas soot, seen previously in K/T sediments, could be transported over long distances in the atmosphere and could have been produced from fossil fuels at the impact site, only burnt vegetation can generate charcoal.

The researchers took samples along a transect from Colorado to Saskatchewan, the nearest site being 2,300 km from the impact site at Chicxulub in Mexico. They found that the amount of charcoal in the K/T-boundary layers was actually several times less than that in other Cretaceous sediments, arguing strongly that there were no impact-induced wildfires across North America. This in turn implies that some previous estimates of the thermal power delivered by the K/T meteorite are substantially too high. **Philip Ball**

Cancer

Sporadic insights

Cell **115**, 523–535 (2003)

Inherited mutations in the *BRCA1* or *BRCA2* genes that result in complete loss of function of the BRCA proteins account for most cases of familial breast and ovarian cancer. But the overwhelming majority of breast and ovarian cancers are sporadic, not inherited, and here mutations in *BRCA* genes are hardly ever seen. Luke Hughes-Davies and colleagues now identify *EMSY* as a gene that is multiplied in such cancers and provides a link to *BRCA2*.

The BRCA proteins are implicated in cellular processes such as regulating gene transcription, repairing DNA and remodelling chromosome architecture, although their precise role is unclear. Hughes-Davies *et al.* find that the EMSY protein can bind BRCA2, silencing its ability to activate transcription. But EMSY also represses transcription directly, localizes to sites of repair following DNA damage, and associates with proteins that regulate chromosome structure. BRCA2 and EMSY seem, physically and functionally, to be intimately linked.

The authors' clinical studies show that *EMSY* is frequently, and almost exclusively, multiplied in sporadic breast

and ovarian cancers. And patients with 'sporadic' *EMSY* amplification or 'familial' *BRCA2* deletion have remarkably similar cancer types.

Furthermore, in patients without metastases, *EMSY* amplification is associated with a poor prognosis.

Hughes-Davies *et al.* propose that they have uncovered a missing link to the BRCA2 pathway in sporadic breast cancer. **Marie-Thérèse Heemels**

Virology

Smallpox selective potential

Proc. Natl Acad. Sci. USA

doi:10.1073/pnas.2435085100 (2003)

Around 10% of Europeans have a degree of protection against HIV because they lack a gene for a receptor, CCR5, that is found on the surface of certain immune cells targeted by the virus. But why does the mutation exist? HIV has not been killing people for long enough for natural selection to have favoured it, and the CCR5 deletion is very rare in people of non-European origin. Alison P. Galvani and Montgomery Slatkin consider that past selection by smallpox could be the answer.

They argue that, like HIV, susceptibility to infection by smallpox virus may have hinged on the presence of CCR5. Even slight protection from smallpox, which kills 30% of those it infects, would have been a strong selective force. Pan-European smallpox pandemics date back at least 2,000 years, enough to account for the incidence of the CCR5 deletion.

An earlier hypothesis pointed to bubonic plague as the selective force. But Galvani and Montgomery conclude that plague did not kill enough Europeans of non-reproductive age, over a long enough period, to be the explanation. However, there have been no studies to identify a mechanism by which a loss of CCR5 might protect against smallpox. As the authors point out, without such data the smallpox hypothesis remains speculative.

Tom Clarke

Vision

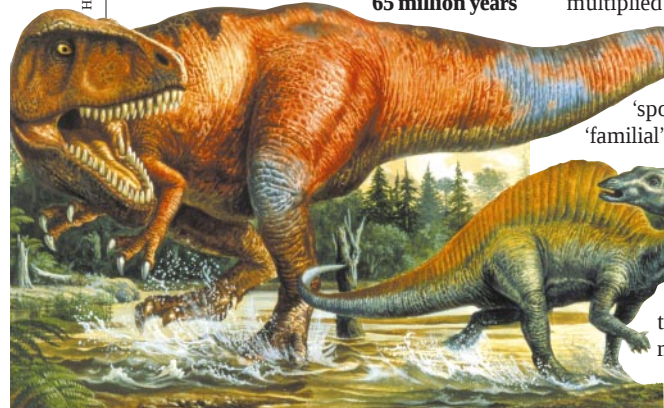
An eye for colour

J. Neurosci. **23**, 10873–10878 (2003)

Flies and other invertebrates can sense light of many different wavelengths, ranging from ultraviolet (UV) to the visible colours. UV vision is used for foraging, navigation and mate selection. One molecule in particular underlies sensitivity to both UV and visible light. Rhodopsin, as it is known, is made of two parts: a light-sensitive chemical group that is bound to the opsin protein.

Ernesto Salcedo *et al.* report that if a specific amino acid in the binding pocket on opsin is a glutamate or asparagine, flies can sense blue light. If the amino acid is a lysine, the invertebrate eye responds to UV.

Previous studies had shown that the same site is altered in some forms of human night blindness, and also mediates UV vision in birds. Amino-acid changes at different sites are responsible for UV vision in fish, reptiles and rodents. Salcedo *et al.* conclude that UV vision may have arisen independently in vertebrate and invertebrate species, but that the molecular mechanism underpinning this type of vision is probably similar. **Helen R. Pilcher**



Solitary wave behaviour of sand dunes

Colliding dunes appear to traverse through one another and emerge unscathed.

Barchan sand dunes are highly mobile, crescent-shaped dunes that occur in areas where sand is sparse and the wind is unidirectional. Here we show mathematically how two such loose-grained dunes are able to pass through one another while still preserving their shape. The crucial parameters for this solitary-wave behaviour, which is consistent with field observations, are the heights of the two colliding dunes.

Because of their remarkable mobility, barchan dunes can disrupt structures such as roads and pipelines in arid, sandy regions. The velocity of a dune can reach several tens of a metres per year and is proportional to the reciprocal of its height¹, meaning that small dunes move faster than large ones^{1,2}. However, the dynamics and evolution of dunes are difficult to assess because of the large time scales involved, with field measurements having to be compiled over several decades^{2–4}.

Several models have therefore attempted to describe dune morphology and formation^{5–10}, for example in terms of the turbulent wind field, saltation sand flux¹ over the windward side, and avalanches travelling down the slip face. Calculations have focused on single dunes or on dune patterns, but not on dune interactions. We now solve the equations of motion that describe what happens when a small barchan dune bumps into a larger one. The results indicate that dunes can behave like solitons¹¹ under certain cir-



Figure 1 A field of crescent-shaped barchan sand dunes in the desert between Chimbote and Casma on the coast of Peru.

cumstances, crossing through one another without changing their shape. (Solitons conform to solutions of nonlinear equations, for example those that describe waves propagating in shallow water¹¹.)

This behaviour is consistent with the observed presence of small barchans on the downwind side of large ones¹², which seemingly indicates that they must have passed through one another without being altered. Small dunes are also found on the lee side of bigger ones in large barchan fields near Laâyoune, Morocco, and in the desert of La Pampa in Peru, for example (Fig. 1). The belief remains widespread, however, that a small barchan will be completely absorbed when it hits a bigger one, in view of the fact that a sand formation cannot cross the slip face of a dune without being destroyed.

We solved a set of equations^{7,8} describing a large heap of sand with the shape of a gaussian function, initially placed downwind of a smaller heap. The strength of the wind blowing into the system was fixed to a shear velocity of 0.5 m s^{-1} . After some time, the gaussians develop into a shape that is typical of barchan dunes¹⁰. The smaller barchan at some point bumps into the larger one, leading to a hybrid state in which the two dunes are fused in a complex pattern. Three different situations can be observed: coalescence, breeding and solitary-wave behaviour, depending on the relative sizes of the two dunes. As control parameters, we chose the relative difference in height between the two dunes, $\Delta h/h_2$, and the height of the smaller dune, h_2 .

The stability of the slip face of the upwind dune crucially influences the final state of the dunes leaving the hybrid state, as well as their relative velocities. When $\Delta h/h_2$ is small, the dunes move with similar velocities and solitary-wave behaviour occurs (Fig. 2). In the intermediate hybrid state, the dune behind is not sufficiently fast-moving to wander all of the way up to the slip face of the bigger dune in front, because it gains so much sand that at some point it becomes larger, and therefore slower, than the one in front. The dune that was previously bigger now becomes the smaller one, and its velocity becomes sufficiently large to leave the hybrid state. Effectively, it seems that the smaller dune crosses the bigger one, whereas in reality the two heaps never merge, owing to mass exchange.

For some values of h_2 and $\Delta h/h_2$, the emerging dune is larger than the incoming one; for other values, it is smaller. This means that they do not behave exactly as solitons, but rather like solitary waves. Intermediate values of h_2 and $\Delta h/h_2$ exist for which the two dunes exactly maintain their sizes and volumes — that is, they behave as solitons.

For smaller $\Delta h/h_2$, we find two different situations. If the height difference between the two dunes is very large, the small dune is entirely swallowed. For moderate height differences, we observe 'breeding' — the creation of two baby dunes at the horns of a barchan.

In regions where the wind is unidirectional and sand is abundantly available, another type of dune forms: these are known as transverse dunes and are translationally invariant. Solitary-wave behaviour should also be possible for transverse dunes. The interaction of laterally shifted dunes needs to be investigated and the behaviour of entire dune fields simulated.

Veit Schwämmle*, Hans J. Herrmann†**

*ICA-1, University of Stuttgart,
70569 Stuttgart, Germany

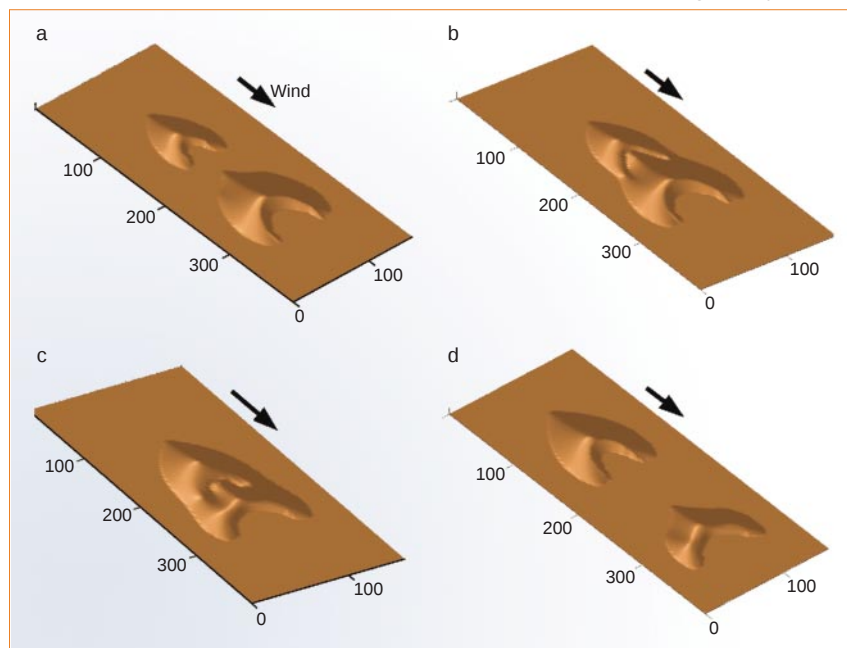


Figure 2 Time series of the solitary-wave behaviour of two barchan sand dunes placed one in front of the other. Parameters are $h_2 = 7.5 \text{ m}$ and $\Delta h/h_2 = 0.9$; distances are in metres. **a**, The dunes in their characteristic forms; **b**, 0.48 years after **a**, the smaller dune bumps into the larger one; **c**, hybrid state 0.63 years after **a**; **d**, the two dunes depart from the hybrid state (1.42 years after **a**).

e-mail: veit@ica1.uni-stuttgart.de

†Departamento de Física, Universidade Federal do Ceará, 60455-970 Fortaleza, Brazil

1. Bagnold, R. A. *The Physics of Blown Sand and Desert Dunes* (Methuen, London, 1941).
2. Finkel, H. J. *J. Geol.* **67**, 614–647 (1959).
3. Hesp, P. A. & Hastings, K. *Geomorphology* **22**, 193–204 (1998).
4. Sauermann, G., Rogon, P., Poliakov, A. & Herrmann, H. J. *Geomorphology* **36**, 47–62 (2000).
5. Wippermann, F. K. & Gross, G. *Boundary Layer Meteorol.* **36**, 319–334 (1986).
6. Werner, B. T. *J. Geol.* **98**, 1–17 (1990).
7. Sauermann, G., Kroy, K. & Herrmann, H. J. *Phys. Rev. E* **64**, 31305 (2001).
8. Kroy, K., Sauermann, G. & Herrmann, H. J. *Phys. Rev. L.* **68**, 54301 (2002).
9. Hersen, P. et al. <http://arxiv.org/abs/cond-mat/0303115> (2003).
10. Schwämmle, V. & Herrmann, H. Preprint <http://arxiv.org/abs/cond-mat/0304695> (2003).
11. Lamb, G. L. *Elements of Soliton Theory* (Wiley, New York, 1980).
12. Besler, H. Z. *Geomorphol. N.F.* **126**, 59–74 (2002).

Competing financial interests: declared none.

Genetics

Influence of TOR kinase on lifespan in *C. elegans*

The group of enzymes known as TOR (for 'target of rapamycin') kinases regulates cell growth and proliferation in response to nutrients and hormone-dependent mitogenic signals^{1,2}. Here we show that TOR deficiency in the nematode *Caenorhabditis elegans* more than doubles its natural lifespan. This new function for TOR signalling in ageing control may represent a link between nutrition, metabolism and longevity.

In *C. elegans*, the absence of LET-363/TOR activity causes developmental arrest at the L3 larval stage³. We examined nematodes bred as *let-363/CeTor* genetic null mutants and nematodes that had been depleted of TOR by using RNA interference to block *let-363* expression (termed *let-363*-(RNAi) worms), and found that these animals had a strikingly extended mean lifespan (Fig. 1a, squares and triangles, respectively). At 25.5 °C, the mean lifetime was 25 days in *let-363* mutants compared with 10 days in wild-type animals. This is all the more intriguing in light of the fact that TOR-deficient worms existed as arrested L3 larvae. In comparison, L3 larval arrest induced by starvation persisted for only 14 days on average in wild-type animals (Fig. 1a, diamonds).

Strong inhibition of mitochondrial respiration also arrests development at the L3 stage, whereas weaker inhibition permits growth to adulthood and extends adult lifespan, but only if it occurs during larval development⁴. In contrast, treatment with *let-363* double-stranded RNA starting from the first day of adulthood lengthens lifespan to a comparable extent when RNAi treatment is initiated at hatching (Fig. 1a, open triangles). This indicates that TOR has a role in ageing control during adulthood and that the long-lived phenotype of *let-363*-(RNAi) adults cannot be explained by reduced mitochondrial activity.

Longevity in *C. elegans* is controlled hor-

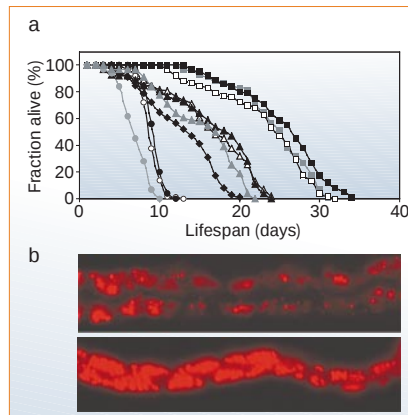


Figure 1 TOR deficiency in the nematode *Caenorhabditis elegans*. **a**, Lifespan of TOR-deficient worms compared with the wild type at 25.5 °C: wild type (filled circles); *daf-16(mg50)* (open circles); *dpy-5(e61) unc-13(e450)* double mutant (shaded circles); TOR-deficient triple mutants *let-363(h114) dpy-5(e61) unc-13(e450)* (open squares); *let-363(h111) dpy-5(e61) unc-13(e450)* (filled squares) and *let-363(h131) dpy-5(e61) unc-13(e450)* (shaded squares); *let-363*-(RNAi)-treated worms from hatching (open triangles) or from the first day of adulthood (shaded triangles); *let-363*-(RNAi)-treated *daf-16(mg50)* worms (filled triangles); starving-arrested wild-type L3 larvae (filled diamonds). Disruption of TOR by RNAi (triangles) seems to be incomplete, as lifespan is not extended as much as in *let-363* mutants (squares). **b**, Nile Red staining of lipid droplets in a wild-type L3 larva (top) and an L3 larva arrested by *let-363*-(RNAi) treatment (bottom). Images were obtained with the same exposure time.

monally by a conserved signalling pathway that involves insulin and insulin-like growth factor (IGF)^{5,6}. Mutants with reduced DAF-2/IGF signalling activity live twice as long as the wild type^{5,6}. The DAF-2/IGF cascade also acts during adulthood to influence ageing⁷. The remarkable similarity in the developmental stage at which ageing rate is affected, and our finding that the extended lifespan of *daf-2(e1370)* mutants is not increased further by treatment with *let-363* RNAi (results not shown) — as it is with RNAi blocking expression of respiratory-chain components⁴ — raise the possibility that TOR and the DAF-2/IGF pathway are related in controlling lifespan.

This idea is compatible with results indicating that the insulin/IGF cascade regulates protein synthesis and cell growth in mammals and *Drosophila* through the activity of nutrient-sensing TOR (reviewed in refs 1, 2, 8). We have also noted that *let-363*-(RNAi) animals share certain features of the pleiotropic *Daf-2*(–) phenotype, such as lipid accumulation mainly in intestinal cells⁹ (Fig. 1b), as well as reduced fertility¹⁰ (mean brood sizes: *let-363*-(RNAi) adults, 68 ± 6.4; wild type, 191 ± 14.5) and reduced viability¹⁰ (embryonic/early larval arrest: *let-363*-(RNAi), 40.3%; wild type, 5.4%).

Strong mutations in DAF-2/IGF signalling cause a long-lived phenotype, together with a state of developmental diapause known as dauer that is triggered by starvation and crowding in the wild type¹¹. According to our results (not shown), *let-363(h111)* animals

bearing the thermosensitive *daf-2(e1370ts)* mutation were able to form dauers at the restrictive temperature. Furthermore, *let-363*-(RNAi) enhanced dauer formation in *daf-2(e1370)* animals. At 20 °C, only 4.6% (29 out of 630) of *daf-2(e1370)* mutants entered into the dauer stage, compared with 17.9% (146 out of 817) of *daf-2(e1370); let-363*-(RNAi) animals (results not shown). This indicates a genetic interaction between *let-363/CeTor* and *daf-2*. These results show that in *C. elegans* the TOR and DAF-2/IGF signalling pathways could be related in controlling ageing, metabolism and reproductive growth.

Lifespan extension in *daf-2(e1370)* mutants requires the activity of the forkhead transcription factor DAF-16 (refs 5, 6). Mutations in *daf-16*, however, do not suppress the long-lived phenotype of *let-363*-(RNAi) worms (Fig. 1a, filled triangles), indicating that TOR may be acting downstream or independently of DAF-16, and that it is interacting with the insulin endocrine system. Although the detailed signalling connections require clarification, our findings point to TOR as a possible mediator of lifespan regulation by insulin signalling and nutrient sensing.

Tibor Vellai*, Krisztina Takacs-Vellai*, Yue Zhang*, Attila L. Kovacs†, László Orosz‡§, Fritz Müller*

*Department of Biology, University of Fribourg, Fribourg 1700, Switzerland

e-mail: fritz.mueller@unifr.ch

†Department of General Zoology and ‡Department of Genetics, Eötvös Loránd University, Budapest 1117, Hungary

§Agricultural Biotechnology Centre, Gödöllő 2100, Hungary

1. Abraham, R. T. *Cell* **111**, 9–12 (2002).
2. Oldham, S. & Hafen, E. *Trends Cell Biol.* **13**, 79–85 (2003).
3. Long, X. et al. *Curr. Biol.* **12**, 1448–1461 (2002).
4. Dillin, A. et al. *Science* **298**, 2398–2401 (2002).
5. Kenyon, C. et al. *Nature* **366**, 461–464 (1993).
6. Kenyon, C. in *C. elegans II* (eds Riddle, D. L., Blumenthal, T., Meyer, B. J. & Priess, J. R.) 791–814 (Cold Spring Harbor Lab. Press, Cold Spring Harbor, New York, 1997).
7. Dillin, A. et al. *Science* **298**, 830–834 (2002).
8. Saltiel, A. R. & Kahn, C. R. *Nature* **414**, 799–806 (2001).
9. Ashrafi, K. et al. *Nature* **421**, 268–272 (2003).
10. Tissenbaum, H. A. & Ruvkun, G. *Genetics* **148**, 703–717 (1998).
11. Riddle, D. L. & Albert, P. S. in *C. elegans II* (eds Riddle, D. L. et al.) 739–768 (Cold Spring Harbor Lab. Press, Cold Spring Harbor, New York, 1997).

Competing financial interests: declared none.

Metabolism

Ecology shapes bird bioenergetics

The basal rate of metabolism of birds and mammals is the lowest rate that is compatible with endothermic temperature regulation, balancing the heat generated with the heat lost by the product of thermal conductance and the temperature differential with the environment¹. Here I measure the bioenergetics of 13 species and 9 genera of birds of paradise (Paradisaeidae) and

show that 99% of the variation in their basal rates of metabolism can be accounted for by interspecific variation in body mass, food habits and altitudinal distribution. These findings, which are derived from 31% of the species and 53% of the genera in this family, give the most complete picture of the standard energetics of any diverse family of birds.

Basal rate, regarded as the standard rate of metabolism in endotherms, conforms to a power function that increases with body mass and varies with food habits, climate, body composition, a continental or island distribution, and possibly other factors¹. Individual and population growth rates², reproductive rates³ and energy expenditure in the field, at least in mammals^{1,4–7}, correlate with variation in basal rate.

Among passerine birds, one of the most distinctive families is the Paradisaeidae, which is found principally in New Guinea. This family is noted for the elaborate plumage of most males, widespread use of lek polygyny, extended longevity, low reproductive rate and diverse food habits, which vary from strict frugivory to extensive insectivory⁸.

In an analysis of covariance, the $\log_{10}$ basal rates of 13 species of Paradisaeidae, measured in rainforest habitat at the Technological University in Lae, Papua New Guinea, simultaneously correlated with $\log_{10}$ body mass ($F_{1,8} = 755.09$, $P < 0.0001$), food habits ($F_{2,8} = 28.78$, $P = 0.0002$) and the altitude at which these species live ($F_{1,8} = 7.98$, $P = 0.023$). Predicted basal rates of paradisaeids (in ml O₂ per hour) are described by

$$V_{O_2}^Y = 3.35(FA)g^{0.879} \quad (1)$$

where F is a non-dimensional coefficient for food habits, A is a non-dimensional coefficient for altitude, and g is body mass in grams. The coefficients are antilogarithms of the intercepts of the curve of $\log_{10}$ basal rate on $\log_{10}$ body mass.

In this case, F is 1.00 for omnivores ($n = 5$), 0.906 for insectivores ($n = 3$) and 0.757 for frugivores ($n = 5$); A is 1.00 for species that live at altitudes greater than 1,000 m ($n = 9$) and 0.906 for species restricted to lower altitudes ($n = 4$). Equation (1) accounts for 99.0% of the variation in basal rate in this family ($F_{1,11} = 1,114.24$, $P < 0.0001$), which assures that all studied species conform to a single relationship, irrespective of body mass, food habits or the altitude at which they live (Fig. 1).

In birds of paradise that preferentially eat fruit and live at altitudes of less than 1,000 m, the combined coefficient in equation (1) is $3.35 \times 0.757 \times 0.906 = 2.30$, whereas in high-altitude omnivores it is $3.35 \times 1.00 \times 1.00 = 3.35$, or 1.46 times that of the low-altitude frugivores. Fruit-eating specialists have basal rates that are 79.4% of the basal rates of species that eat insects as more than 10% of their diet, and species that

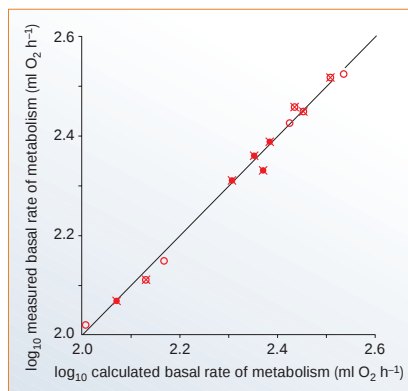


Figure 1 $\log_{10}$ basal rates of metabolism measured in birds of paradise as a function of $\log_{10}$ basal metabolic rates predicted by equation (1). Food habits: hollow circles, insect and mixed diet; filled circles, fruit diet. Crosses indicate birds living at altitudes above 1,000 metres.

live at altitudes below 1,000 m have basal rates that are 90.6% of those that live at higher altitudes. Other combinations of body mass, food habits and altitude account for the observed basal rates of the remaining species. Body mass alone accounts for 91.7% of the variation in basal rate; 97.5% of this variation is accounted for by the combination of body mass and food habits.

A low basal metabolic rate in frugivorous birds is also seen in manakins (Pipridae)⁹, pigeons (Columbidae)¹⁰, toucans (Ramphastidae)¹¹ and a hornbill (Bucerotidae)¹¹, although frugivorous passerines have higher basal rates than frugivorous non-passerines, contrary to the view that there is no difference in the basal rates of non-passerines and passerines¹². An examination of the evolutionary relationships within the Paradisaeidae indicates that frugivory¹³ and low basal rate are plesiomorphic conditions in this family. The evolution of high basal rates in the studied species occurred at least three times in relation to movement into higher altitudes, and twice in relation to the adoption of an omnivorous or insectivorous diet. The use of phylogeny as the 'cause' for the evolution of character states^{7,12} is therefore doubtful.

The remarkable plumage of males in dimorphic species of this family seem to entail very high costs during synthesis: rates of metabolism increase by at least 53% during moulting in male king birds of paradise (*Cicinnurus regius*). However, brightly coloured males in species with sexually dimorphic plumage do not have higher basal rates than dull-coloured females of the same species ($F_{2,4} = 0.46$, $P = 0.66$), and sexually dimorphic species do not have higher basal rates than sexually monomorphic species ($F_{1,8} = 0.87$, $P = 0.37$). Furthermore, species that use lek displays do not have higher basal rates ($F_{1,10} = 4.45$, $P = 0.079$) than those that use solitary displays.

The behaviour and ecology of the Paradisaeidae has been extensively studied, but the

energetics of these species have not been evaluated until now. The analysis of covariance that I used for this task is superior to the conventional phylogenetic-contrasts approach.

Brian K. McNab

Department of Zoology, University of Florida,
Gainesville, Florida 32611, USA

e-mail: bkm@zoo.ufl.edu

1. McNab, B. K. *The Physiological Ecology of Vertebrates: A View from Energetics* (Cornell Univ. Press, Ithaca, New York, 2002).
2. McNab, B. K. *Am. Nat.* **116**, 106–124 (1980).
3. Glazier, D. S. *Comp. Biochem. Physiol.* **80** A, 587–590 (1985).
4. Nagy, K. A. *Ecol. Monogr.* **57**, 111–128 (1987).
5. Nagy, K. A. *Aust. J. Zool.* **42**, 43–53 (1994).
6. Koteja, P. *Funct. Ecol.* **5**, 56–64 (1991).
7. Ricklefs, R. E., Konarzewski, M. & Daan, S. *Am. Nat.* **147**, 1047–1071 (1996).
8. Frith, C. B. & Beehler, B. M. *The Birds of Paradise* (Oxford Univ. Press, Oxford, 1998).
9. Bartholomew, G. A., Vleck, C. M. & Bucher, T. L. *Physiol. Zool.* **56**, 370–379 (1983).
10. McNab, B. K. *Comp. Biochem. Physiol.* **127** A, 309–329 (2000).
11. McNab, B. K. *Auk* **118**, 916–933 (2001).
12. Reynolds, P. S. & Lee, R. M. *Am. Nat.* **147**, 735–759 (1996).
13. Beehler, B. M. in *Acta XIX Congr. Internat. Ornithol.* (ed. Ouellet, H.) 816–828 (Ottawa Univ. Press, Ottawa, 1989).

Competing financial interests: declared none.

COMMUNICATIONS ARISING

Geochronology

Dating of the Herto hominin fossils

The age of Pleistocene *Homo sapiens* fossils and archaeological material from the Bouri Formation in the Middle Awash region of Ethiopia, discovered in 1997 by White and colleagues^{1,2}, has been constrained to between 160 ± 2 and 154 ± 7 kyr on the basis of isotopic dating and stratigraphic and geochemical evidence². However, our analysis of their stratigraphic and geochronological data indicate that, although the estimated maximum age (160 ± 2 kyr) is valid, the minimum age (154 ± 7 kyr) is doubtful. These important discoveries³ may therefore be distinctly younger than reported^{1,2}.

The fossils and archaeological remains occur in volcanoclastic sandstones and gravel deposits of the older part of the Upper Herto Member of the Bouri Formation. The lower boundary of this member is formed by a widespread erosional surface. Immediately below this surface is a bentonite tuff (MA97-1, 2) which has been isotopically dated to 260 ± 16 kyr. Two pumices and two obsidian clasts from the fossiliferous deposits above the erosional surface provide a relatively narrow age range close to 160 kyr, which indicates a maximum age for the deposit, but it cannot necessarily be inferred that it is "close to the actual time of deposition of this fossiliferous unit"².

The main problem in dating the fossiliferous horizon is how to establish a younger age limit. The Upper Herto Member is capped by the stratigraphically important Waideo Vitric Tuff (WAVT, MA92-1). Unfortunately, this has proved difficult to date isotopically because of contamination by older feldspar

phenocrysts, so the only available constraint on the age of the deposit is the maximum age inferred from the volcanic clasts.

To establish a minimum age, the major elements and 11 trace elements (Zn, Cu, Cr, Zr, V, Sc, Nb, Ba, Sr, Rb and Y) of glass shards of the Waidedo Vitric Tuff were geochemically analysed and used for correlation with other, previously dated Pleistocene tuffs — Clark *et al.* refer to an unnamed tuff from the Konso region of southern Ethiopia, more than 500 km away. Although this tuff is geochemically comparable to the Waidedo Vitric Tuff, it had not itself been dated either; however, the age of the immediately overlying Konso Silver Tuff (TG 120), 154 ± 7 kyr, was used as the younger age limit for the fossiliferous deposits of Middle Awash.

However, this “geochemical correlation”, which is methodically not comparable to a real stratigraphic correlation, is highly speculative. Isotope investigations and rare-earth-element analysis are both needed before the correlation can be accepted with confidence. The recognized geochemical similarities do not mean that the two tuffs from Middle Awash and southern Ethiopia definitely belong to the same volcanic eruptive event — they simply indicate similar petrogenetic conditions and are not necessarily of the same age.

It is even difficult to find a particular eruptive centre on the basis of such limited data, because many Quaternary volcanic centres of similar silicic composition are located within the Main Ethiopian Rift between the Afar Depression and southern Ethiopia⁴. Several volcanic deposits of roughly similar composition, but of different ages, should therefore be expected in the Pleistocene succession of the whole Main Ethiopian Rift.

We contend that the narrow age range of the *H. sapiens* fossils and archaeological remains from the Upper Herto Member of the Middle Awash, as estimated by Clark *et al.*, is too optimistic. A real minimum age can be established only if a volcanic horizon can be dated directly in or above the fossiliferous succession. Until then, only the maximum age of 160 kyr can reliably be used for this important anthropological material, although the fossils and archaeological remains could be considerably younger than this.

Peter Faupl, Wolfram Richter, Christoph Urbanek

Department of Geological Sciences, University of Vienna, 1090 Vienna, Austria

e-mail: peter.faupl@univie.ac.at

- White, T. D. *et al.* *Nature* **423**, 742–747 (2003).
- Clark, J. D. *et al.* *Nature* **423**, 747–752 (2003).
- Stringer, C. *Nature* **423**, 692–695 (2003).
- Chernet, T. *et al.* *J. Volcan. Geotherm. Res.* **80**, 267–280 (1998).

Clark *et al.* reply — Faupl *et al.* question our confidence in the geochemical correlation that anchors the younger age constraint of 154 kyr on the Herto antiquities¹. According to convention, we reported the tephrochem-

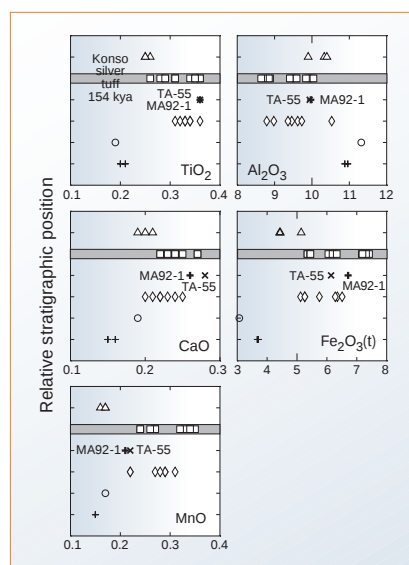


Figure 1 Average results from electron-probe microanalyses of distinct glass populations from multimodal Konso-region tephra glass, compared with unimodal Konso tephra TA-55, the correlative Upper Herto Waidedo Vitric Tuff (MA92-1) and older unimodal Konso tephra. Tephra are placed in relative stratigraphic context on the basis of field observations in the Konso region. Major elements are recalculated to 100% anhydrous and are plotted as percentage by weight of oxide on the x-axes. Symbols are average analyses of other Konso tephra layers; the same symbols are used to show the glass populations from a given tephra layer.

ical correlation as the most probable explanation for the available data. Contrary to published information^{2,3}, Faupl *et al.* contend that Quaternary volcanic centres in the Main Ethiopian Rift and the Afar depression are of “similar silicic composition”. They provide no datum and no specific counter-argument to our correlation. Moreover, the single citation⁴ on which they base their assertion that our correlation is “highly speculative” contains no geochemical data.

The following facts bear on the contested correlation: the WAVT lies stratigraphically above a 160-kyr datum, whereas the Konso tuff is the first volcanic horizon stratigraphically below a 154-kyr datum; the WAVT and the Konso tuff are both fine vitric ashes with sparse, fragmented crystals of quartz and predominantly accidental feldspar derived from older crustal reservoirs (as shown by Ar–Ar geochronology), and homogeneous, bubble-wall and pumiceous glass shards. Analysis of individual glass shards from the WAVT and the Konso tuff yielded chemically homogeneous single glass populations with a coefficient of similarity^{5,6} of 0.95; comparison of WAVT and the Konso tuff to regional tephra from major silicic volcanoes (6 Myr to Recent) in the Main Ethiopian and Afar Rifts, including the Middle Awash ($n > 300$ samples) and Konso ($n = 39$), yielded no correlation.

Furthermore, analysis of purified glass separates from the WAVT and the Konso tuff yielded data on major-element oxides and

trace elements that confirm this correlation. Combined bulk glass major- and trace-element data yielded a coefficient of similarity of 0.93. Our unpublished analysis of nine rare-earth elements (La, Ce, Nd, Sm, Gd, Dy, Er, Yb and Lu) in the same bulk glass separates yields a coefficient of similarity of 0.97, which further substantiates the published correlation¹.

Figure 1 illustrates the relative stratigraphic position of the Konso tuff (TA-55) with respect to other identified Pleistocene tephra. Tephra positioned stratigraphically above and below the Konso Tuff, including the 154-kyr Konso Silver Tuff, contain heterogeneous and/or multimodal glass populations. This contrasts with the homogeneous, unimodal glass population observed in the Konso tuff and the correlative Upper Herto WAVT (MA92-1). Moreover, none of the individual glass populations in these other Pleistocene tephra yields chemical characteristics that are as similar to either TA-55 or MA92-1, as observed between these two tephra (Fig. 1).

The eruption that produced the WAVT, which is more than 2 m thick and is preserved in the Upper Herto Member (MA92-1) and in the Konso region (TA-55), was comparable in volume to large eruptions that produced other ash-fall deposits that have been tephrostratigraphically and tephrochronologically correlated between sites in the Middle Awash, the Turkana Basin and the Gulf of Aden^{7,8}. Although the identification of the specific volcanic sources of these large eruptions would be desirable, it is not a prerequisite for tephrochemical correlation.

Field observations in the Konso region place the TA-55 marker horizon stratigraphically below the Konso Silver Tuff dated at 154 kyr. This, as well as the identification of this marker horizon stratigraphically above the 160-kyr fossil-bearing sands of the Upper Herto Member of the Bouri Formation, Middle Awash, Ethiopia, provides convincing evidence that the Upper Herto archaeological and palaeontological remains, including the newly identified *Homo sapiens idaltu*, are securely constrained to be between 160 ± 2 and 154 ± 7 kyr old, as we stated¹.

William K. Hart, Giday WoldeGabriel, Shigehiro Katoh, Paul R. Renne, Gen Suwa, Berhane Asfaw, Tim D. White*

*Department of Integrative Biology and Laboratory for Human Evolutionary Studies, Museum of Vertebrate Zoology, University of California, Berkeley, California 94720, USA

e-mail: timwhite@socrates.berkeley.edu

- Clark, J. D. *et al.* *Nature* **423**, 747–752 (2003).
- WoldeGabriel, G. *et al.* *Acta Vulcanologica* **11**, 53–67 (1999).
- WoldeGabriel, G. *et al.* *Quat. Internat.* **13/14**, 69–76 (1992).
- Chernet, T. *et al.* *J. Volcanol. Geotherm. Res.* **80**, 267–280 (1998).
- Borchardt, G. A., Aruscavage, P. J. & Millard, H. T. Jr *J. Sed. Pet.* **42**, 301–306 (1972).
- Sarna-Wojcicki, A. M. in *Quaternary Geochronology: Methods and Applications* (eds Noller, J. S., Sowers, J. M. & Lettis, W. R.) 357–377 (AGU Ref. Shelf 4, Washington, DC, 2000).
- Brown, F. H. *Nature* **300**, 631–633 (1982).
- Brown, F. H. *et al.* *Quat. Internat.* **13/14**, 55–67 (1992).

Centre-surround inhibition among olfactory bulb glomeruli

J. L. Aungst^{1*}, P. M. Heyward^{1*†}, A. C. Puche^{1*}, S. V. Karnup¹, A. Hayar¹, G. Szabo² & M. T. Shipley¹

¹Department of Anatomy and Neurobiology, Program in Neuroscience, The University of Maryland School of Medicine, Room 222, 685 West Baltimore Street, Baltimore, Maryland 21201, USA

²Department of Gene Technology and Developmental Neurobiology, Institute of Experimental Medicine, Budapest, Hungary

* These authors contributed equally to this work

† Present address: Department of Physiology, Otago School of Medical Sciences, University of Otago, Dunedin, 9001, New Zealand

Centre-surround inhibition—the suppression of activity of neighbouring cells by a central group of neurons—is a fundamental mechanism that increases contrast in patterned sensory processing. The initial stage of neural processing in olfaction occurs in olfactory bulb glomeruli, but evidence for functional interactions between glomeruli is fragmentary. Here we show that the so-called ‘short axon’ cells, contrary to their name, send interglomerular axons over long distances to form excitatory synapses with inhibitory periglomerular neurons up to 20–30 glomeruli away. Interglomerular excitation of these periglomerular cells potentially inhibits mitral cells and forms an on-centre, off-surround circuit. This interglomerular centre-surround inhibitory network, along with the well-established mitral-granule-mitral inhibitory circuit, forms a serial, two-stage inhibitory circuit that could enhance spatiotemporal responses to odours.

Odour perception begins with the binding of odorant molecules to specific receptors expressed by olfactory receptor neurons (ORN). Each ORN expresses a single odorant receptor from a repertoire of ~1,000 genes¹. ORN axons project to the main olfactory bulb (MOB), where they synapse onto the dendrites of second-order neurons in spheroidal structures—glomeruli. Individual odorant receptors are broadly tuned: a given receptor is activated by many odours and, conversely, an odour activates many different receptors. As each glomerulus is innervated by ORNs expressing the same odorant receptor^{1–5}, different odours produce unique patterns of multiglomerular activation^{6–14}. Therefore, odour perception requires the neural computation of patterns of glomerular activity.

Current hypotheses of MOB function hold that lateral interactions between mitral/tufted output cells, mediated by inhibitory granule cells, are fundamental to the processing of spatiotemporal patterns of glomerular activity^{15,16}. Strongly activated mitral/tufted cells inhibit weakly activated ones through dendrodendritic synapses with granule cells. This is thought to enhance discrimination at the level of MOB output. Lateral interactions may also occur at the level of olfactory nerve input through connections between glomeruli¹⁷. Although interglomerular connections have been suggested by classical Golgi and electron microscopy (EM) studies^{18,19}, the magnitude and spatial extent of these connections are unknown, and physiological evidence for interglomerular interactions is lacking. The goal of the present study, therefore, was to determine the anatomical basis for and functional significance of interglomerular connections.

Extensive interglomerular connections

To investigate the anatomical basis of interglomerular projections, small tracer injections were made into the glomerular layer of the intact mouse MOB to label juxtaglomerular (JG) cells that give rise to interglomerular connections (Fig. 1 and Supplementary Fig. 1). A single subglomerular-sized (5–10 µm diameter) injection of DiI (5–10 µm diameter) was made into each MOB (Fig. 1a). Analysis of labelled JG cell distributions (48.5 ± 8 cells per MOB, *n* = 6; Fig. 1b–e) showed that 50% of labelled cells were located at distances greater than 350 µm (~5–7 glomeruli) from the injection site, and 10% were at distances greater than 850 µm (~15–18 glomeruli). As all injection sites were smaller than the diameter of a single glomerulus

(50–100 µm), these results underestimate the number of cells projecting to a single glomerulus, but do show that interglomerular connections extend over much greater distances than was previously thought. Labelled JG cells were distributed radially with no significant rostrocaudal or dorsoventral anisotropy (Fig. 1c, d). Identical tracing experiments in rats showed that interglomerular projections were organized as in mice (75.3 ± 8.8 cells per MOB, *n* = 3 rats; Fig. 1e); 50% of cells were located more than 600 µm (~6 or 7 glomeruli) from the injection site, and 10% more than 1,000 µm away (~10–12 glomeruli). When corrected for bulb size (glomerular layer surface area), there were no significant differences in cell distribution between rat and mouse.

As DiI is taken up by all membranes and undergoes both anterograde and retrograde transport, rhodamine microbeads—a retrograde tracer taken up preferentially at synaptic endings²⁰—were injected into the glomerular layer *in vivo* to confirm our anatomical results. The distribution of microbead-labelled cells from a single injection in each animal was similar to that seen with DiI (44.5 ± 13.7 cells, *n* = 4 mice; Fig. 1e, f), suggesting a widespread axonal, rather than dendritic, projection from these cells.

To investigate the trajectories of interglomerular axons, we used a novel *in vitro* parasagittal surface slice of the medial wall of the MOB (Fig. 2a). This ‘surface slice’ contained the olfactory nerve layer (ONL), the glomerular layer and the superficial external plexiform layer (EPL), but not the mitral cell layer and granule cell layer. This surface slice provided an ‘aerial view’ of most medial glomeruli. A single subglomerular-sized (5–10 µm diameter, *n* = 16 slices) iontophoretic DiI injection into each slice labelled a dense plexus of interglomerular connections (Fig. 2b). Interglomerular axons radiated in all directions with the highest process density within 500–600 µm of the injection site (Fig. 2b). Some processes were labelled as far as 2 mm from the injection site. This dense interglomerular network arose from only a small number of JG cells (59.5 ± 12.3 cells per slice), indicating that the cell type(s) interconnecting glomeruli have numerous axons and/or extensive axonal branching.

SA cells form interglomerular connections

Taken together, the tract-tracing results show that connections between olfactory bulb glomeruli are more substantial and extend

over far greater distances than was previously appreciated, but do not indicate which JG cell type(s) gives rise to this network. There are three principal types of JG neuron: periglomerular (PG), external tufted (ET) and short axon (SA) cells^{17,18,19}. It has generally been thought that interglomerular connections are inhibitory and derive from GABA (γ -aminobutyric acid)-synthesizing PG cells¹⁶, but there is little direct evidence for this. To investigate the role of GABAergic neurons in interglomerular projections, we made Dil injections in transgenic mice in which the promoter for glutamic acid decarboxylase 65 (GAD65, the principal GABA-synthesizing enzyme in the MOB²¹) drives the expression of green fluorescent protein (GFP; Fig. 2c). Dil injections into the glomerular layer of these mice (51 ± 15.8 cells per bulb, $n = 3$ mice; Fig. 1e) labelled

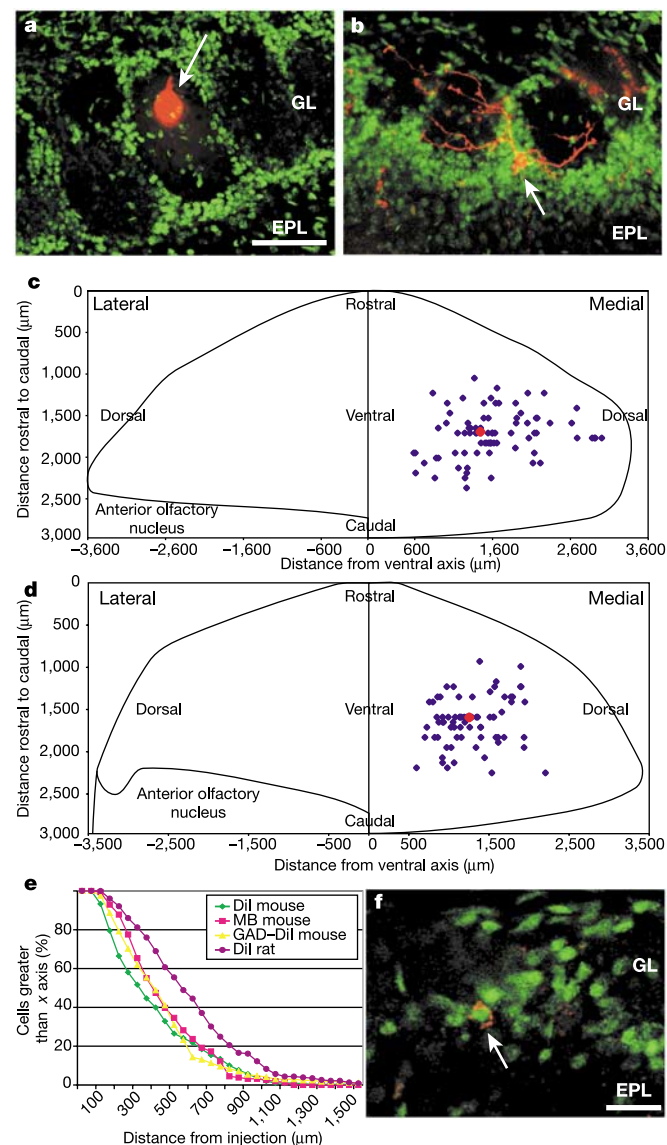


Figure 1 Extensive interglomerular connections shown by Dil and microbead tracing. **a**, Subglomerular-sized iontophoretic Dil injection into a single glomerulus (arrow). GL, glomerular layer. **b**, Dil-labelled JG cell bodies (arrowhead) within the glomerular layer (nuclei counterstained green). **c**, **d**, 'Flat plane' representations of the olfactory bulb glomerular layer from two experiments, showing the distribution of all labelled JG cells (cell bodies, blue dots; injection site, red dot). **e**, Percentage of labelled cells plotted as a function of distance from the injection site. Distributions in the larger rat bulb are significantly greater than those in the mouse ($P < 0.001$). MB, rhodamine microbeads. **f**, Microbead-labelled JG cells (arrows). Scale bar in **a** represents 50 μm and applies to **a** and **b**; scale bar in **f** represents 20 μm .

cells with a distribution identical to that of wild-type mice. However, only $9.1 \pm 1.8\%$ of Dil-labelled cells located one glomerulus or more from the injected glomerulus contained GFP (Fig. 2d and Supplementary Fig. 2). Therefore, the contribution of

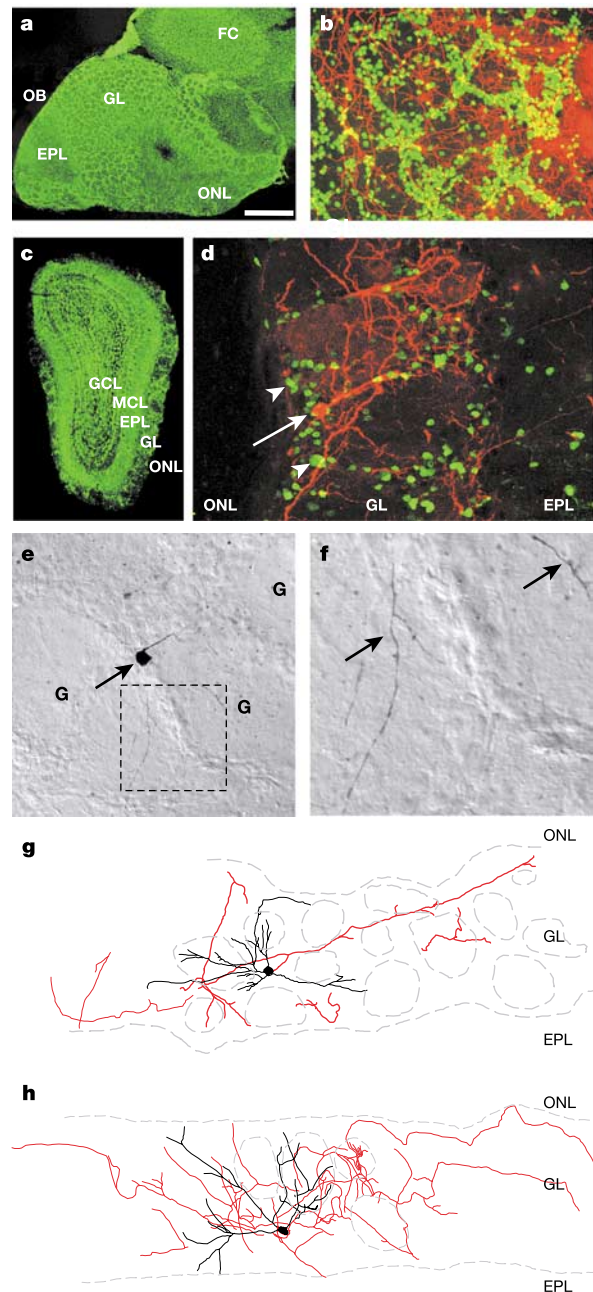


Figure 2 Interglomerular connections derive from non-GABAergic 'short axon' cells. **a**, Surface slice showing the medial glomeruli stained with a nuclear counterstain (green). Owing to curvature of the bulb, areas of ONL, glomerular layer and EPL are visible in the same section. FC, frontal cortex; OB, olfactory bulb. **b**, Dil-labelled interglomerular connections form a dense plexus (injection off panel to right). **c**, Robust green fluorescence in GABAergic neurons in GAD65-GFP mice. GCL, granule cell layer; MCL, mitral cell layer. **d**, A subpopulation of PG cells expresses GFP (arrowheads). More than 90% of interglomerular projection cells lacked GFP (the arrow shows a Dil⁺ GFP⁻ cell 350 μm from the injection site). **e**, Biocytin-filled SA cell soma situated between two glomeruli under DIC (differential interference contrast) optics. G, glomerulus. **f**, Demarcated region in **e** showing the thin SA cell axon extending through two glomeruli (arrows). **g**, **h**, Reconstruction of two biocytin-filled SA cells (dendrites, black; axon, red). Scale bar in **a** represents 1 mm in **a**; 100 μm in **b**; 500 μm in **c**; 25 μm in **d**, **e**; 10 μm in **f**; and 80 μm in **g**, **h**.

GABAergic cells to interglomerular connections is minor. Furthermore, none of the labelled cells had prominent dendritic tufts characteristic of DiI-labelled ET cells, suggesting that ET cells do not contribute substantially to the interglomerular network.

We next analysed a population of over 200 physiologically characterized, biocytin-filled JG neurons available from other experiments. Neither PG nor ET cells had interglomerular axons projecting over the distances revealed by our tracing experiments. Only one cell type consistently had long axons within the glomerular layer. This neuron had relatively thin axons ($<0.5\ \mu\text{m}$ diameter, $n = 11$ cells; Fig. 2e, f). These axons ramified up to $850\ \mu\text{m}$ exclusively within the glomerular layer, with no obvious directional bias (Fig. 2g, h; total axon length, $3,625 \pm 816\ \mu\text{m}$). The cells had 3–5 (3.2 ± 0.3) poorly branched dendrites that contacted 2–4 glomeruli (total dendrite length, $1,308 \pm 64\ \mu\text{m}$). With the notable exception of the length and richness of their axonal trajectories, the morphological features of these cells correspond to the ‘short axon’ cell described in classical Golgi studies. Classical SA cells were said to have 3–5 ‘stout’, infrequently branching dendrites that traverse 2–4 glomeruli, and a thin varicose axon projecting over 1–2 glomeruli^{17,22}. The greater extent of the axonal arborizations in biocytin-filled SA cells versus the ‘short axons’ seen in classical studies is

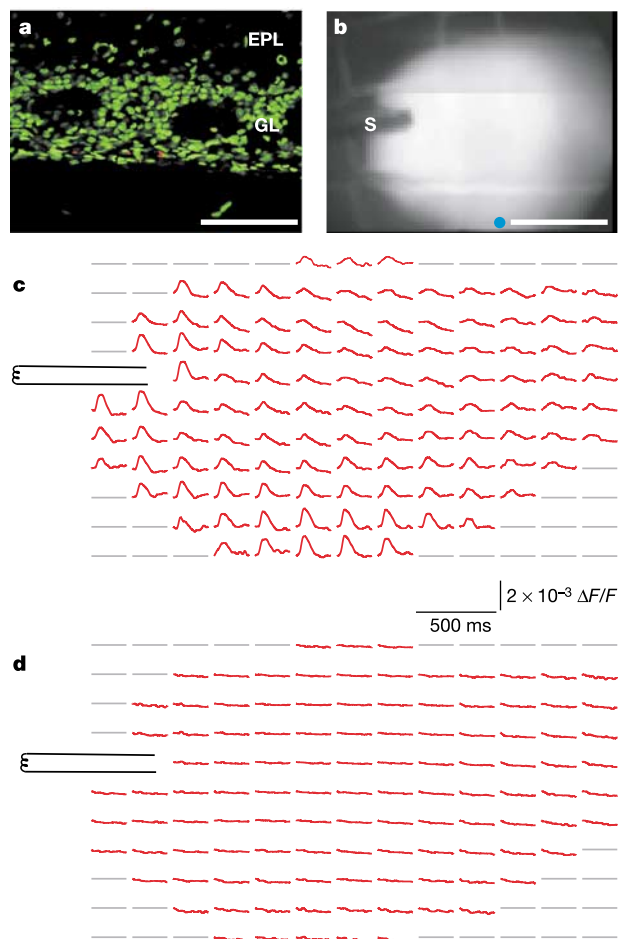


Figure 3 Functional imaging shows extensive interglomerular excitation. **a**, DiI tracing confirming the elimination of olfactory nerve axons (red) in the glomerular layer of a mouse after ablation by ZnSO_4 lavage. **b**, Fluorescence image of a parasagittal surface slice loaded *in vitro* with the voltage-sensitive dye RH414. The cyan circle shows the approximate diameter of a single mouse glomerulus. **s**, stimulating electrode. **c**, Spatial averages ($100 \times 100\ \mu\text{m}$) of optical signals evoked by glomerular stimulation. Stimulation induces widespread depolarization (upward deflection in the inverted traces). **d**, Responses were abolished when glutamatergic transmission was blocked. Scale bars represent $100\ \mu\text{m}$ in **a** and $500\ \mu\text{m}$ in **b**.

probably due to limited impregnation of axons achieved with the Golgi–Cox method^{17,22}. Accordingly, we conclude that the main source of interglomerular connections is the SA cell.

Interglomerular connections are excitatory

Our anatomical findings demonstrate an extensive interglomerular network but provide no insight into its functional significance. To investigate the physiology of this system, we first used functional imaging. As the anatomical results indicate that interglomerular connections are not GABAergic, we reasoned that interglomerular connections might be glutamatergic. If true, then focal glomerular stimulation should, through the interglomerular network, cause widespread excitation of JG cells. To test this hypothesis, we used a voltage-sensitive dye to monitor changes in membrane potential after focal stimulation of the glomerular layer in a surface slice (Fig. 3). To prevent activation of olfactory nerve axons, the slices were taken from mice in which the olfactory nerve had been irreversibly ablated by ZnSO_4 lavage²³ of the olfactory mucosa 8–12 days before recording (Fig. 3a).

Focal glomerular stimulation ($n = 6$ slices) evoked widespread depolarization throughout the glomerular layer (Fig. 3c). The depolarizing response spread in all directions from the stimulating electrode, extending $900 \pm 93\ \mu\text{m}$ along the long axis of the MOB (rise time, $51.3 \pm 4.7\ \text{ms}$; half-width, $51.4 \pm 8.3\ \text{ms}$). These widespread depolarizing responses were abolished by the addition of the ionotropic glutamate receptor antagonists APV (D-2-amino-5-phosphonopentanoate; $50\ \mu\text{M}$) and CNQX (6-cyano-7-nitro-quinoline-2,3-dione; $10\ \mu\text{M}$) ($n = 3$; Fig. 3d). Responses recovered after 40 min of washout. These results show that interglomerular connections are extensive, excitatory and mediated by glutamate acting at ionotropic receptors.

Although the VSD (voltage-sensitive dye) imaging results indicate that excitatory synapses underlie a substantial proportion of interglomerular connections, they do not exclude some contribution of long-range inhibitory connections. To further investigate

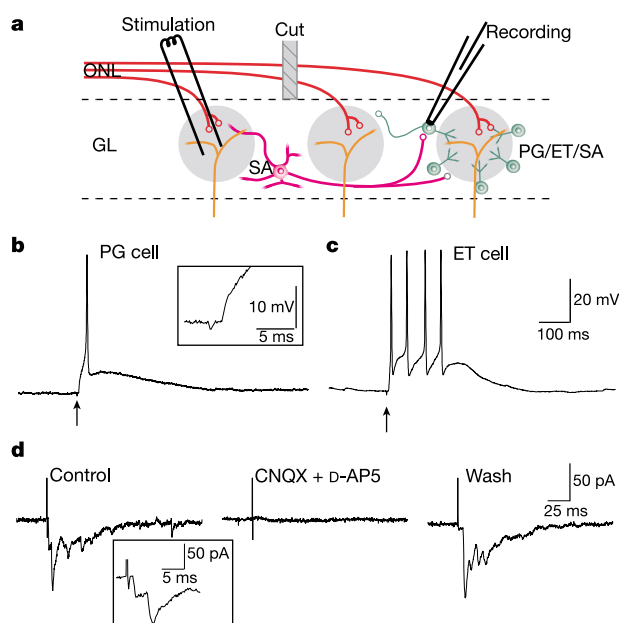


Figure 4 Interglomerular connections excite juxtglomerular cells. **a**, Schematic of the experimental design. **b**, **c**, PG cells (**b**) and ET cells (**c**) responded with EPSPs to distant glomerular layer stimulation (arrow). The inset (expanded) shows a 2 ms response latency, consistent with a monosynaptic event. **d**, Voltage-clamp recordings of an EPSC in a PG cell following distant glomerular stimulation (left). The inset (expanded) shows 1.8 ms EPSC response latency. The EPSC was abolished when glutamatergic transmission was blocked (middle), and recovered after washout (right).

the circuit actions of interglomerular connections, whole-cell recordings were made from JG cells 3 or 4 glomeruli away from the site of focal glomerular stimulation in MOB slices. The ONL was severed between the stimulation and recording sites to prevent propagation through olfactory nerve fibres (Fig. 4a). Fourteen biocytin-filled JG cells that responded to glomerular layer stimulation were classified according to established morphological criteria^{19,22,24}: 10 (72%) were PG cells, 3 (21%) were ET cells, and 1 (7%) was an SA cell. Most of these cells (93%, 13 of 14) responded with excitatory postsynaptic potentials (EPSPs) after distant glomerular layer stimulation (Fig. 4b, c), whereas only one PG cell responded with an inhibitory postsynaptic potential (IPSP). Response latencies were 2.1 ± 0.2 ms for EPSPs, and 1.8 ms for the single case of an IPSP, consistent with monosynaptic events. Evoked EPSPs were abolished in the presence of ionotropic glutamate receptor antagonists (APV and CNQX; $n = 6$; Fig. 4d). These results further support the conclusion that most interglomerular connections are excitatory and glutamatergic.

Centre-surround inhibition among glomeruli

Taken together, these results show that SA cells form an extensive network of excitatory interglomerular connections that release

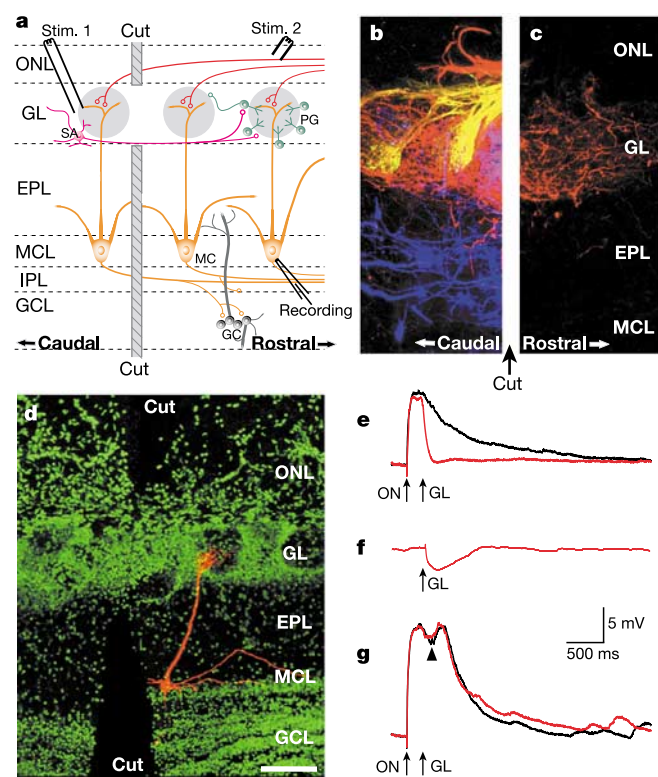


Figure 5 Interglomerular connections inhibit mitral cell responses to olfactory nerve input. **a**, Schematic of the experimental design. **b**, Microinjections of DiA, Dil and DiD into the ONL, glomerular layer and EPL, respectively, caudal to the cuts label processes coursing tangentially through intact regions of the slice. GC, granule cell; IPL, internal plexiform layer; MC, mitral cell. **c**, On the opposite side of the cuts (rostral), only interglomerular connections are intact. **d**, Biocytin-filled mitral cell (red) and nuclear counterstain (green). The cuts through the ONL and deep bulb layers are clearly visible. **e**, Mitral cells respond to ONL stimulation with an LLD (black trace) that is abruptly terminated by distant glomerular layer stimulation (arrow, red trace). **f**, Stimulation of interglomerular connections evokes an IPSP in mitral cells. **g**, In the presence of gabazine, olfactory-nerve-evoked LLDs are no longer suppressed by glomerular layer stimulation (black trace, olfactory nerve stimulation alone; red trace, olfactory nerve plus interglomerular stimulation). Consistent with blockade of inhibition, the LLD is larger and has an additional peak (arrowhead). The scale bar in **d** represents $100 \mu\text{m}$ in **b–d**.

glutamate and excite PG cells. PG cells are thought to form GABAergic²⁵ inhibitory synapses with the dendrites of mitral cells. Therefore, activation of interglomerular connections should inhibit mitral cell responses to olfactory nerve input. To test this hypothesis, we recorded mitral cell responses to olfactory nerve stimulation in MOB slices²⁶ that had been surgically altered to allow interglomerular connections to be stimulated in isolation. Microsurgical cuts were made through the ONL and all deep layers, leaving only the

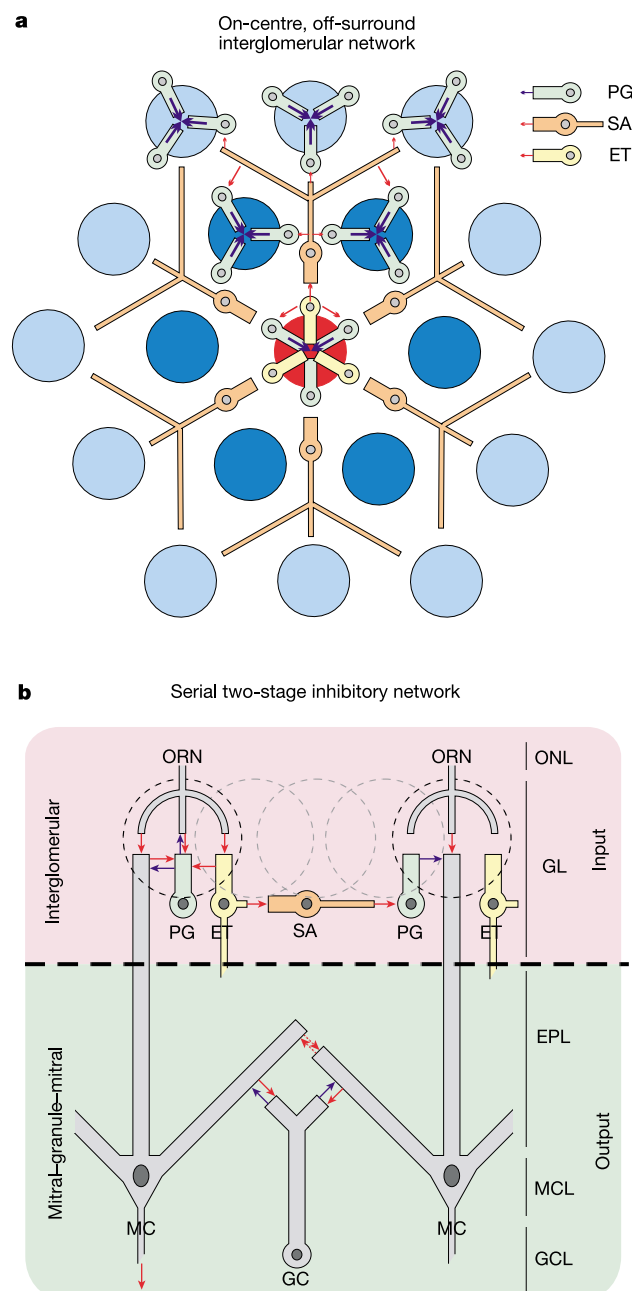


Figure 6 Centre-surround inhibitory networks in the olfactory bulb. **a**, Schematic of the on-centre, off-surround inhibitory network in the glomerular layer (red arrows, excitatory connections; blue arrows, inhibitory). An activated glomerulus (red) causes widespread inhibition (blue) of surrounding glomeruli by the interglomerular network. **b**, Schematic showing two serial levels of MOB centre-surround inhibition: interglomerular and mitral-granule-mitral. The interglomerular circuit at the level of olfactory nerve input is formed by SA cells that excite PG and ET cells in distant glomeruli. Activation of GABAergic PG cells causes inhibition of mitral cells. The mitral-granule-mitral network at the level of mitral cell output is formed by mitral cells exciting granule cells, which in turn inhibit the same (feedback) and other (feedforward) mitral cells.

glomerular layer intact between opposite sides of the cuts (Fig. 5a–d). The olfactory nerve was stimulated and mitral cell responses recorded on one side of the cuts while interglomerular connections were stimulated on the other side. After recording, we made microinjections of DiA, DiI and DiD into the ONL, glomerular layer and EPL, respectively. This tracing showed that only interglomerular connections were intact (Fig. 5b, c). Biocytin fills of recorded mitral cells showed that processes on the side of the cut were severed (Fig. 5d) and that the apical dendrite was intact.

Mitral cells respond to olfactory nerve stimulation with a long-lasting depolarization (LLD)²⁷. Activation of interglomerular connections excites GABAergic PG cells and should, therefore, inhibit evoked LLDs in mitral cells. Indeed, stimulation of interglomerular connections abruptly terminated olfactory-nerve-evoked LLDs in 6 of 7 mitral cells recorded 150–525 μ m distal to glomerular layer stimulation (Fig. 5e). Glomerular stimulation reduced LLD magnitude, calculated by integrated voltage under the curve, to 0.15% of control ($n = 6$, $P < 0.01$). Thus, input from distant regions of the glomerular layer inhibits mitral cell responses to olfactory nerve input. Mean latency from olfactory nerve stimulation to LLD onset—a monosynaptic event—was 2.9 ± 0.2 ms, and mean latency from interglomerular stimulation to the onset of LLD termination was 6.5 ± 1.2 ms. This latency suggests that interglomerular connections inhibit mitral cells through monosynaptic activation of GABAergic PG cells or disynaptic activation of ET cells that monosynaptically excite PG cells, or both.

Interglomerular excitation of PG cells should thus generate a summated IPSP in glomerular tufts of mitral cells sufficient to terminate the LLD. Detection of IPSPs in mitral cell glomerular tufts is limited by remoteness from the somatic recording site. To enhance detection of distal IPSPs, mitral cells were held 30 mV positive to the Cl^- equilibrium potential (at -55 to -50 mV, $E_{\text{Cl}} = -84$ mV). In 3 of 4 mitral cells in which LLDs were terminated by interglomerular stimulation, an IPSP was detectable (mean response latency, 5.7 ± 0.9 ms, $n = 3$; Fig. 5f). As other potential sources of inhibitory input were excluded by microsurgical cuts, these IPSPs must be generated in the apical tuft of the mitral cell by interglomerular activation of GABAergic PG cells.

If interglomerular connections act through PG cells to attenuate the LLD, this action should be abolished by blockade of GABAergic transmission. Application of the GABA_A receptor blocker gabazine (5 μ M) increased LLD magnitude and introduced additional peaks in the response (Fig. 5g), consistent with the suppression of tonic GABAergic inhibition. Importantly, in the presence of gabazine, stimulation of interglomerular connections no longer attenuated olfactory-nerve-evoked LLDs (Fig. 5g). The effect of gabazine reversed after 40 min of washout. These experiments show that excitatory interglomerular connections activate local GABAergic PG cells to cause postsynaptic inhibition of mitral cell responses to olfactory nerve input.

Discussion

Centre-surround inhibition is a fundamental mechanism in sensory processing²⁸ that enhances contrast between populations of strongly and weakly excited neurons. The best-characterized inhibitory circuit in the MOB involves lateral inhibition of mitral cells at mitral-granule cell synapses in the EPL^{16,29,30}. By comparison, virtually nothing is known about centre-surround inhibition in the glomerular layer, the initial site of synaptic processing in the olfactory system. The findings presented here demonstrate the presence of long-range interconnections between glomeruli, extending up to 2 mm (20–30 glomeruli radii). These connections excite PG cells in distant glomeruli, causing inhibition of mitral cell responses to olfactory nerve input. This interglomerular network provides on-centre, off-surround processing at the initial stage of synaptic integration in the MOB (Fig. 6a). By enhancing spatial contrast between patterns of odour-evoked glomerular activity at

the level of sensory input, centre-surround processing may be a key first step in the neural computation of odours.

Our findings show that SA cells are the principal JG neuron type making interglomerular connections. The classical picture of interglomerular connections was based on the Golgi–Cox staining studies of Cajal and Golgi^{22,24}, and the EM, Golgi and degeneration studies of Pinching and Powell^{17–19,31}. The axonal projections of SA cells were reported to extend only 1–3 glomeruli from the cell soma^{19,22}. However, as the Golgi–Cox method stains axons incompletely^{17,22}, the classical view of the SA cell underestimated the extent of its axonal arborization. Indeed, our DiI and microbead tracer experiments, and reconstructions of individual biocytin-filled neurons, showed that SA cell axons extend several millimetres within the glomerular layer.

Imaging studies of olfactory nerve terminals loaded with calcium-sensitive dyes, which reflect only presynaptic olfactory nerve terminal activity, show that individual or small groups of glomeruli are activated by low concentrations of single odorants^{32,33}. By contrast, studies with 2-deoxyglucose (2-DG)^{9,34–36}, c-fos^{37,38} and functional magnetic resonance imaging (fMRI)¹³, which reflect the activation of pre- and/or postsynaptic elements, show broader regions of activity spanning multiple glomeruli. The interglomerular connections reported here may reconcile these disparate observations. As activation of a single glomerulus excites JG cells hundreds of micrometres away, methods such as 2-DG, c-fos and fMRI may reflect initial activation of a few discrete glomeruli by a single odorant followed by synaptic activation of JG cells in multiple neighbouring glomeruli through the interglomerular circuit.

Interglomerular connections synapse not only onto PG cells but also onto ET cells and other SA cells. ET cells excite PG and SA cells associated with the same glomerulus. Martinez and Freeman observed an excitatory field potential at the glomerular layer–EPL border, thought to reflect activity in mitral³⁹ and/or tufted cells. Interglomerular excitation of SA and ET cells may underlie these excitatory field potentials. In the intact bulb, sequential excitation of ET and SA cells could propagate activity over very long distances in the glomerular layer.

On-centre, off-surround organization in the MOB has generally been attributed to synaptic interactions between the lateral dendrites of mitral cells and inhibitory granule cells in the EPL^{15,16}. In this mitral-granule-mitral circuit, strongly-activated mitral cells⁴⁰ laterally inhibit weakly-activated surrounding mitral cells. The interglomerular circuit reported here also forms a centre-surround organization: activation of a glomerulus—‘on-centre’—should cause widespread inhibition of mitral cells in neighbouring glomeruli—‘off-surround’. Thus, the current findings add a new dimension to our understanding of the functional organization of the olfactory bulb by showing that there are two serial stages of centre-surround inhibition (Fig. 6b). Neural network modelling suggests that two serial stages of centre-surround inhibition are highly advantageous to sensory networks^{41,42}. The first stage is the interglomerular circuit, which operates at the level of olfactory nerve sensory input. It is formed by SA cells that make excitatory interglomerular connections onto inhibitory PG cells, which in turn inhibit mitral cells. This stage also provides presynaptic regulation of transmitter release from olfactory nerve terminals^{32,43}. This stage may preserve relative response magnitudes across a range of input intensities (pattern normalization), reduce noise (low-band filtration) and enhance contrast between activated glomeruli. The second stage operates at the level of mitral cell output. It is formed by the mitral-granule-mitral circuit. The second stage may prevent response saturation at high input intensities and provide further contrast enhancement⁴². Together, these two centre-surround circuits may synergistically enhance patterns of olfactory input.

The neural representation of odour refined by these two stages of inhibition need not be static, and probably involves spatiotemporal changes in the distribution of activity in the MOB⁴⁴. In this view, the

mitral–granule–mitral circuit serves to progressively sharpen odour representations in the neuronal population, rather than simply sharpening the tuning curves of individual mitral cells. However, because the mitral–granule–mitral circuit will amplify differences in patterns of glomerular activity⁴⁴, pattern normalization and effective noise reduction is needed at the first stage of input to the olfactory bulb, operations that may be performed by the centre-surround interglomerular network. □

Methods

Tracer injections

For whole-brain staining with DiI, adult (5-week-old) mice (CD-1 and transgenic GAD65–GFP; the generation of these mice will be published elsewhere) and rats (Sprague–Dawley) were anaesthetized (150 mg kg^{−1} Nembutal) and transcardially perfused with 4% paraformaldehyde in 0.1 M phosphate buffer (PFA; pH 7.4). Their brains were removed and a single glomerulus in each MOB injected iontophoretically with DiI (1 mg ml^{−1} in ethanol, 0.5–2.0 μ A, tip diameter 1.5 μ m). Brains were incubated at 23 °C in PFA for 2 weeks, serially sectioned (60 μ m in the coronal plane), incubated in 20 nM Sytox Green (Molecular Probes) in PBS for 30 min (CD1 and rat) and coverslipped with a DABCO-based mounting media.

For the DiI staining of parasagittal surface slices, brains from adult mice were removed, hemisected and positioned in low-gel agarose with the medial surface parallel to the plane of the vibratome. A single 350 μ m section was cut from the medial surface of each MOB and immobilized in a recording chamber perfused with artificial cerebral spinal fluid (ACSF) containing 120 mM NaCl, 8 mM KCl, 1.3 mM CaCl₂, 1.3 mM MgSO₄, 10 mM glucose, 25 mM NaHCO₃ and 5 mM BES, and saturated with 95% O₂/5% CO₂. This ‘surface slice’ contained the entire medial wall ONL, glomerular layer and part of the EPL. A single glomerulus was iontophoretically injected with DiI, incubated at 23 °C in PFA for 2 weeks, resectioned (50 μ m in the sagittal plane) and stained with Sytox.

For whole-brain staining with microbeads, adult CD-1 mice were anaesthetized and 10–20 nl of rhodamine-labelled latex microbeads (Lumafuor; 1:2 dilution in saline) were injected into the medial glomerular layer of the left MOB (stereotaxic coordinates 4.28 mm Bregma, 1.02 mm depth, 0.12 mm lateral). The injection pipette was positioned 4.8 mm lateral at 55° over the right MOB, and inserted through the right MOB into the left MOB glomerular layer (1.25 mm from the right MOB surface). After one week, mice were transcardially perfused, and the brain was serially sectioned (40 μ m in the coronal plane) and stained with Sytox.

Sections were visualized at high magnification and the position of labelled JG cell somata plotted on a camera lucida outline. The rostrocaudal distance (*z* axis) of each cell was measured from the section number, and the dorsoventral (*x* axis) and mediolateral (*y* axis) coordinates automatically measured using the Image Processing Toolkit (Reinder Graphics). These coordinates were used to generate reverse cumulative probability graphs of cell distance from the injection using a custom Excel worksheet. Statistical significance was tested by analysis of variance (ANOVA; Scheffe’s post hoc test; all values presented as mean $\pm$ s.e.m.). The surface of each MOB was presented as an unfolded sheet—a ‘flat plane’ representation—by tracing the glomerular layer in each section and ‘unfolding’ these traces into a flattened sheet.

SA cell reconstructions

SA cells were recorded in horizontal rat MOB slices (400 μ m) as previously described²⁶. The pipette filling solution contained 0.5% biocytin, 125 mM potassium gluconate, 2 mM MgCl₂, 10 mM HEPES, 2 mM Mg₂ATP, 0.2 mM Na₃GTP, 1 mM NaCl and 0.2 mM EGTA, pH 7.2. After recording, slices were fixed in PFA, resectioned (80 μ m) and reacted with ABC reagent (Vector Laboratories) and diaminobenzidine following standard protocols. Cells were reconstructed using Neurolucida (Microbrightfield).

Optical recordings

Membrane potential changes were visualized using the voltage-sensitive dye RH414 (Molecular Probes)^{45,46}. Surface slices were prepared from 5–6-week-old mice. The olfactory nerve had been irreversibly destroyed by lavage of the epithelium with ZnSO₄ (ref. 23) 8–12 days before slicing. To confirm the extent of this lesion, several mice were transcardially perfused and the nasal cavities flushed with 5 mg ml^{−1} DiI in ethanol, incubated at 30 °C for 2 weeks, sectioned and imaged.

Surface slices were incubated in constant-flow ACSF for 30 min at 30 °C, then 200 μ M RH414 in ACSF for 30 min at 23 °C, and transferred to a recording bath at 23 °C. Images were captured at 2 kHz through a X10 water immersion objective with a NeuroCCD camera and analysed in NeuroPlex (RedShirtImaging). Slices were illuminated with a stabilized 150 W Xenon lamp (Opti-Quip) with exposure regulated by electronic shutter (Uniblitz, Vincent Associates). Constant-current stimuli (100–300 μ A) were applied through a bipolar stimulating electrode.

Electrophysiology

JG cells were recorded in horizontal rat MOB slices with the glomerular layer stimulated (100–300 μ A) 3–4 glomeruli from the recording site; the ONL between stimulation and recording sites was transected perpendicular to the glomerular layer.

To record from mitral cells, horizontal mouse MOB slices (350 μ M) were prepared with opposing cuts made perpendicular to the glomerular layer through the ONL and all deep layers (EPL to granule cell layer), leaving only the glomerular layer intact. Mitral cells on the rostral side of the cuts were recorded in whole-cell current clamp (held at −70 mV to prevent action potentials). The ONL was stimulated rostral to the cuts and

interglomerular connections stimulated with a second electrode positioned in the glomerular layer caudal to the cuts. The olfactory nerve was stimulated with and without subsequent stimulation (200 ms delay) of the distant glomerular layer. This interval allowed the interglomerular circuit to be investigated without involvement of olfactory-nerve-evoked presynaptic effects⁴⁷. Olfactory nerve stimulation intensity was the minimum sufficient to activate a LLD in 100% of trials (80–300 μ A, 0.1 ms, 0.2 Hz). Glomerular layer stimulation intensity was 300–500 μ A. Sets of at least 20 stimuli in each condition were averaged and area under the voltage traces measured. Statistical significance was tested using Student’s paired two-sample *t*-test (all values presented as mean $\pm$ s.e.m.). After recording, biocytin was visualized with streptavidin-Cy3 (Jackson) or the completeness of cuts verified with lipophilic tract tracing. DiA, DiI and DiD were injected into the ONL, glomerular layer and EPL, respectively, and incubated at 30 °C for one week.

Received 2 June; accepted 21 October 2003; doi:10.1038/nature02185.

- Ressler, K. J., Sullivan, S. L. & Buck, L. B. A zonal organization of odorant receptor gene expression in the olfactory epithelium. *Cell* **73**, 597–609 (1993).
- Mombaerts, P. *et al.* The molecular biology of olfactory perception. *Cold Spring Harb. Symp. Quant. Biol.* **61**, 135–145 (1996).
- Vassar, R. *et al.* Topographic organization of sensory projections to the olfactory bulb. *Cell* **79**, 981–991 (1994).
- Potter, S. M. *et al.* Structure and emergence of specific olfactory glomeruli in the mouse. *J. Neurosci.* **21**, 9713–9723 (2001).
- Treloar, H. B., Feinstein, P., Mombaerts, P. & Greer, C. A. Specificity of glomerular targeting by olfactory sensory axons. *J. Neurosci.* **22**, 2469–2477 (2002).
- Greer, C. A., Stewart, W. B., Kauer, J. S. & Shepherd, G. M. Topographical and laminar localization of 2-deoxyglucose uptake in rat olfactory bulb induced by electrical stimulation of olfactory nerves. *Brain Res.* **217**, 279–293 (1981).
- Halasz, N. & Greer, C. A. Terminal arborizations of olfactory nerve fibers in the glomeruli of the olfactory bulb. *J. Comp. Neurol.* **337**, 307–316 (1993).
- Jastreboff, P. J. *et al.* Specific olfactory receptor populations projecting to identified glomeruli in the rat olfactory bulb. *Proc. Natl Acad. Sci. USA* **81**, 5250–5254 (1984).
- Johnson, B. A. & Leon, M. Modular representations of odorants in the glomerular layer of the rat olfactory bulb and the effects of stimulus concentration. *J. Comp. Neurol.* **422**, 496–509 (2000).
- Johnson, B. A. & Leon, M. Spatial distribution of [¹⁴C]2-deoxyglucose uptake in the glomerular layer of the rat olfactory bulb following early odor preference learning. *J. Comp. Neurol.* **376**, 557–566 (1996).
- Rubin, B. D. & Katz, L. C. Optical imaging of odorant representations in the mammalian olfactory bulb. *Neuron* **23**, 499–511 (1999).
- Xu, F., Kida, I., Hyder, F. & Shulman, R. G. Assessment and discrimination of odor stimuli in rat olfactory bulb by dynamic functional MRI. *Proc. Natl Acad. Sci. USA* **97**, 10601–10606 (2000).
- Yang, X. *et al.* Dynamic mapping at the laminar level of odor-elicited responses in rat olfactory bulb by functional MRI. *Proc. Natl Acad. Sci. USA* **95**, 7715–7720 (1998).
- Rubin, B. D. & Katz, L. C. Spatial coding of enantiomers in the rat olfactory bulb. *Nature Neurosci.* **4**, 355–356 (2001).
- Mori, K., Nagao, H. & Yoshihara, Y. The olfactory bulb: coding and processing of odor molecule information. *Science* **286**, 711–715 (1999).
- Yokoi, M., Mori, K. & Nakanishi, S. Refinement of odor molecule tuning by dendrodendritic synaptic inhibition in the olfactory bulb. *Proc. Natl Acad. Sci. USA* **92**, 3371–3375 (1995).
- Pinching, A. J. & Powell, T. P. Experimental studies on the axons intrinsic to the glomerular layer of the olfactory bulb. *J. Cell Sci.* **10**, 637–655 (1972).
- Pinching, A. J. & Powell, T. P. The neuropil of the periglomerular region of the olfactory bulb. *J. Cell Sci.* **9**, 379–409 (1971).
- Pinching, A. J. & Powell, T. P. The neuron types of the glomerular layer of the olfactory bulb. *J. Cell Sci.* **9**, 305–345 (1971).
- Cornwall, J. & Phillipson, O. T. Quantitative analysis of axonal branching using the retrograde transport of fluorescent latex microspheres. *J. Neurosci. Methods* **24**, 1–9 (1988).
- Sheikh, S. N., Martin, S. B. & Martin, D. L. Regional distribution and relative amounts of glutamate decarboxylase isoforms in rat and mouse brain. *Neurochem. Int.* **35**, 73–80 (1999).
- Cajal, R. S. *Histologie du Systeme Nerveux de l’Homme et des Vertebres* (Maloine, Paris, 1911).
- Kawano, T. & Margolis, F. L. Transsynaptic regulation of olfactory bulb catecholamines in mice and rats. *J. Neurochem.* **39**, 342–348 (1982).
- Golgi, C. *Sulla Fina Struttura dei Bulbi Olfattorii* (Reggio-Emilia, Rome, 1875).
- Mugnaini, E., Oertel, W. H. & Wouterlood, F. F. Immunocytochemical localization of GABA neurons and dopamine neurons in the rat main and accessory olfactory bulbs. *Neurosci. Lett.* **47**, 221–226 (1984).
- Heyward, P., Ennis, M., Keller, A. & Shipley, M. T. Membrane bistability in olfactory bulb mitral cells. *J. Neurosci.* **21**, 5311–5320 (2001).
- Carlson, G. C., Shipley, M. T. & Keller, A. Long-lasting depolarizations in mitral cells of the rat olfactory bulb. *J. Neurosci.* **20**, 2011–2021 (2000).
- Shepherd, G. M. *The Synaptic Organization of the Brain* (Oxford Univ. Press, New York, 1990).
- Chen, W. R., Shen, G. Y., Shepherd, G. M., Hines, M. L. & Midgaard, J. Multiple modes of action potential initiation and propagation in mitral cell primary dendrite. *J. Neurophysiol.* **88**, 2755–2764 (2002).
- Chen, W. R., Xiong, W. & Shepherd, G. M. Analysis of relations between NMDA receptors and GABA release at olfactory bulb reciprocal synapses. *Neuron* **25**, 625–633 (2000).
- Pinching, A. J. & Powell, T. P. The neuropil of the glomeruli of the olfactory bulb. *J. Cell Sci.* **9**, 347–377 (1971).
- Wachowiak, M. & Cohen, L. B. Presynaptic inhibition of primary olfactory afferents mediated by different mechanisms in lobster and turtle. *J. Neurosci.* **19**, 8808–8817 (1999).
- Wachowiak, M. & Cohen, L. B. Presynaptic afferent inhibition of lobster olfactory receptor cells: reduced action-potential propagation into axon terminals. *J. Neurophysiol.* **80**, 1011–1015 (1998).
- Astic, L. & Cattarelli, M. Metabolic mapping of functional activity in the rat olfactory system after a bilateral transection of the lateral olfactory tract. *Brain Res.* **245**, 17–25 (1982).

35. Benson, T. E., Burd, G. D., Greer, C. A., Landis, D. M. & Shepherd, G. M. High-resolution 2-deoxyglucose autoradiography in quick-frozen slabs of neonatal rat olfactory bulb. *Brain Res.* **339**, 67–78 (1985).
36. Johnson, B. A., Woo, C. C. & Leon, M. Spatial coding of odorant features in the glomerular layer of the rat olfactory bulb. *J. Comp. Neurol.* **393**, 457–471 (1998).
37. Guthrie, K. M. & Gall, C. M. Functional mapping of odor-activated neurons in the olfactory bulb. *Chem. Senses* **20**, 271–282 (1995).
38. Onoda, N. Odor-induced fos-like immunoreactivity in the rat olfactory bulb. *Neurosci. Lett.* **137**, 157–160 (1992).
39. Martinez, D. P. & Freeman, W. J. Periglomerular cell action on mitral cells in olfactory bulb shown by current source density analysis. *Brain Res.* **308**, 223–233 (1984).
40. Isaacson, J. S. Glutamate spillover mediates excitatory transmission in the rat olfactory bulb. *Neuron* **23**, 377–384 (1999).
41. Grossberg, S. Neural pattern discrimination. *J. Theor. Biol.* **27**, 291–337 (1970).
42. Grossberg, S. Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biol. Cybern.* **23**, 121–134 (1976).
43. Aroniadou-Anderjaska, V., Ennis, M. & Shipley, M. T. Glomerular synaptic responses to olfactory nerve input in rat olfactory bulb slices. *Neuroscience* **79**, 425–434 (1997).
44. Laurent, G. *et al.* Odor encoding as an active, dynamical process: experiments, computation, and theory. *Annu. Rev. Neurosci.* **24**, 263–297 (2001).
45. Kauer, J. S., Senseman, D. M. & Cohen, L. B. Odor-elicited activity monitored simultaneously from 124 regions of the salamander olfactory bulb using a voltage-sensitive dye. *Brain Res.* **418**, 255–261 (1987).
46. Kauer, J. S. Real-time imaging of evoked activity in local circuits of the salamander olfactory bulb. *Nature* **331**, 166–168 (1988).
47. Aroniadou-Anderjaska, V., Zhou, F. M., Priest, C. A., Ennis, M. & Shipley, M. T. Tonic and synaptically evoked presynaptic inhibition of sensory input to the rat olfactory bulb via GABA_B heteroreceptors. *J. Neurophysiol.* **84**, 1194–1203 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements The authors thank F. L. Margolis for expert assistance in ZnSO₄ epithelium lesions and A. Keller for valuable comments on the manuscript. This work was supported by the National Institute on Deafness and Other Communication Disorders.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.T.S. (mshipley@umaryland.edu).

Crystal structure of plant photosystem I

Adam Ben-Shem¹, Felix Frolow² & Nathan Nelson¹

¹Department of Biochemistry and ²Molecular Microbiology and Biotechnology, The George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel

Oxygenic photosynthesis is the principal producer of both oxygen and organic matter on Earth. The conversion of sunlight into chemical energy is driven by two multisubunit membrane protein complexes named photosystem I and II. We determined the crystal structure of the complete photosystem I (PSI) from a higher plant (*Pisum sativum* var. *alaska*) to 4.4 Å resolution. Its intricate structure shows 12 core subunits, 4 different light-harvesting membrane proteins (LHCI) assembled in a half-moon shape on one side of the core, 45 transmembrane helices, 167 chlorophylls, 3 Fe–S clusters and 2 phylloquinones. About 20 chlorophylls are positioned in strategic locations in the cleft between LHCI and the core. This structure provides a framework for exploration not only of energy and electron transfer but also of the evolutionary forces that shaped the photosynthetic apparatus of terrestrial plants after the divergence of chloroplasts from marine cyanobacteria one billion years ago.

Oxygenic photosynthesis, the conversion of sunlight into chemical energy by plants, green algae and cyanobacteria, underpins the survival of virtually all higher life-forms. By producing oxygen and assimilating carbon dioxide into organic matter it determines to a large extent the composition of our atmosphere and provides essential food and fuel. This process is driven by PSI and PSII, two large multisubunit protein complexes that are embedded in the thylakoid membrane and act in series^{1,2}. Photons absorbed by these complexes induce excitation of a special pair of chlorophylls initiating translocation of an electron across the membrane. This leads to the formation of an electrochemical potential, which powers ATP synthesis³. Water is the electron donor for this process and is oxidized to O₂ and four H⁺ ions by PSII. The electrons extracted from water are shuttled through a quinone pool and the *b₆f* complex to plastocyanin, a small soluble copper protein⁴. Solar energy absorbed by PSI induces translocation of an electron from plastocyanin at the inner face of the membrane (lumen) to ferredoxin on the opposite side (stroma). The redox potential of ferredoxin is subsequently used in numerous regulatory cycles and reactions, including nitrate assimilation, fatty acid desaturation and NADPH production. In the dark, CO₂ reduction to carbohydrates is fuelled by ATP and NADPH chemical energy⁵.

Plant PSI is composed of a reaction centre of up to 14 subunits and a membrane-associated antenna complex (LHCI) that captures light and guides its energy to the reaction centre¹. On the whole, plant PSI binds approximately 200 pigments. Despite its complexity, PSI is highly efficient and almost every photon absorbed results in excitation of the special chlorophyll pair P₇₀₀⁶. LHCI consists of four different membrane proteins (Lhca1–4) with varying stoichiometry depending on light intensity and other environmental factors⁷. As all LHCI proteins share high sequence homology and spectral properties, the need for four different genes is not obvious⁸. LHCI proteins are unique among the chlorophyll-*a/b* binding proteins in their red-shifted absorbance and in the formation of dimers⁹.

The cyanobacterial PSI is smaller in size compared with plant PSI, having a reaction centre that resembles the one in plants but with no peripheral antenna. Its structure (from *Synechococcus elongatus*) was recently published, providing detailed insights into the molecular architecture of this complex¹⁰. The model contained 12 protein subunits and 127 cofactors (96 chlorophyll *a*, 22 carotenoids, 2 phylloquinones, 3 Fe₄–S₄ clusters and 4 lipids). We sought to understand the interactions within LHCI and between it and the reaction centre, and to reveal how adjustment to a terrestrial habitat shaped the plant complex after divergence from marine cyanobacteria. To this end, we determined the X-ray crystal structure of PSI from peas and describe here a model at 4.4 Å resolution.

Overall architecture

Plant PSI is monomeric both *in vitro* and *in vivo*^{11,12}. A view from the stroma of the plant PSI C α backbone (Fig. 1a) reveals that the reaction centre and LHCI form two distinct and loosely associated moieties, with a deep cleft between them. The four antenna proteins assemble into two dimers, arranged in a series, creating a half-moon-shaped belt that docks to the reaction centre's subunit F side. We assigned (see Supplementary Information) these two dimers to Lhca1–Lhca4 and Lhca2–Lhca3 (Fig. 1a) according to published biochemical and mutagenic studies^{13–17}. The LHCI belt, with its associated chlorophylls, is the most prominent addition to the PSI structure made by plants (and green algae). It contributes a mass of 150 kDa out of approximately 525 kDa. The plant reaction centre moiety retains the location and orientation of the electron transfer components and all cyanobacterial transmembrane helices, except those of subunits X and M (not present in plants). In addition to these retained features, four reaction centre proteins are present exclusively in plants and green algae (subunits G, H, N, O; see refs 12, 18). Two of these, namely G (PsaG) and H (PsaH), both membrane proteins of 10 kDa, are revealed in our model (Fig. 1a, b). A single transmembrane helix adjacent to PsaL was assigned to PsaH in agreement with cross-linking data¹³. This transmembrane segment is followed by a 20-Å-long helix lying parallel to the membrane, coordinating one chlorophyll molecule (Fig. 2a). The position and shape of PsaH conforms well to its proposed role as a docking site for LHCII¹⁹. On the opposite side of the reaction centre PsaG, with its two tilting transmembrane helices, contributes most of the contact surface area for association with LHCI (Fig. 2a).

On the luminal side, the most noticeable distinction between plant and cyanobacterial reaction centres is the helix–loop–helix motif contributed by the longer amino-terminal domain of plant PsaF (Fig. 1b). This domain enables the more efficient plastocyanin binding in plants and, as a result, two orders of magnitude faster electron transfer from this copper protein to P₇₀₀²⁰.

Conservation of chlorophyll positions in the PSI core

Recent phylogenetic studies and whole-genome analyses depict an evolutionary model in which oxygen-evolving cyanobacteria originated after the divergence of Bacteria from Archaea, becoming prevalent around 2.5 billion years ago, and coinciding with the earliest rise in oxygen levels on Earth^{21,22}. Since chloroplasts diverged from their cyanobacterial ancestors, at least 1 billion years ago, they have undergone a separate evolution. It is therefore remarkable that in the structure of the plant PSI reaction centre, we find that the positions of virtually all of the cyanobacterial chlorophylls are almost precisely preserved (Fig. 2b). During 1 billion years of evolutionary processes an enormous number of random

mutations took place, yet both cyanobacteria and higher plants maintained the same chlorophyll arrangement. Out of the 96 chlorophyll molecules reported in the model of the cyanobacterial PSI reaction centre only three are missing in the plant PSI reaction centre: two bound to PsaM and PsaX, and one bound to PsaJ. From the remaining 93 chlorophylls, 92 are identified at the same position in the plant reaction centre, including 15 chlorophylls with their Mg^{2+} coordinated by water (Fig. 2a, b and ref. 10). It is noteworthy that among the more peripheral chlorophylls of the reaction centre, which probably do not transfer energy directly to P_{700} , we find only one that is not conserved in the plant complex (J3) and only one other where a significant positional modification occurs (B33). A few other minor alterations, mainly in chromophore orientation, are also observed. Thus, adaptation of the reaction centre to the utilization of energy from the LHCI antenna only required the addition of ten chlorophyll molecules at the three 'contact regions', as discussed below (Fig. 2c).

Formation of the LHCI belt and its binding to the core

What are the interactions giving rise to the formation of the two

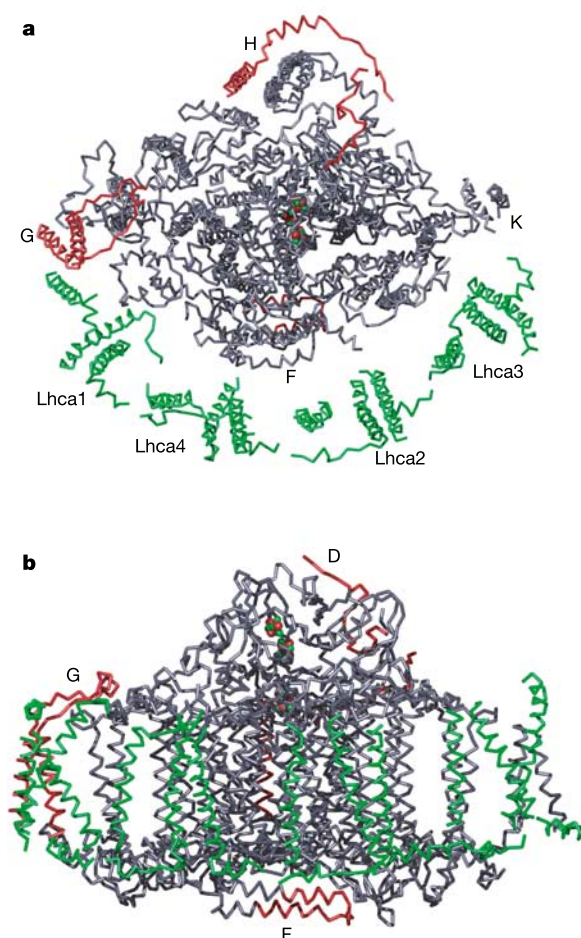


Figure 1 The structural model of plant PSI at 4.4 Å represented as $\text{C}\alpha$ backbone. The four light-harvesting proteins are in green (Lhca1–4). Novel structural elements within the reaction centre (core) that are not present in the cyanobacterial counterpart are coloured red; conserved features of the reaction centre are in grey. The three Fe_4S_4 clusters are depicted as red (Fe) and green (S) balls. **a**, View from the stromal side of the thylakoid membrane. Subunits F, G, H and K of the reaction centre are indicated. The assignment of the four different Lhca proteins is shown. **b**, A view from the LHCI side. Subunits F, G and D are indicated. The helix-loop-helix N-terminal domain of subunit F and the N terminus of subunit D unique to plant photosystem I are coloured red.

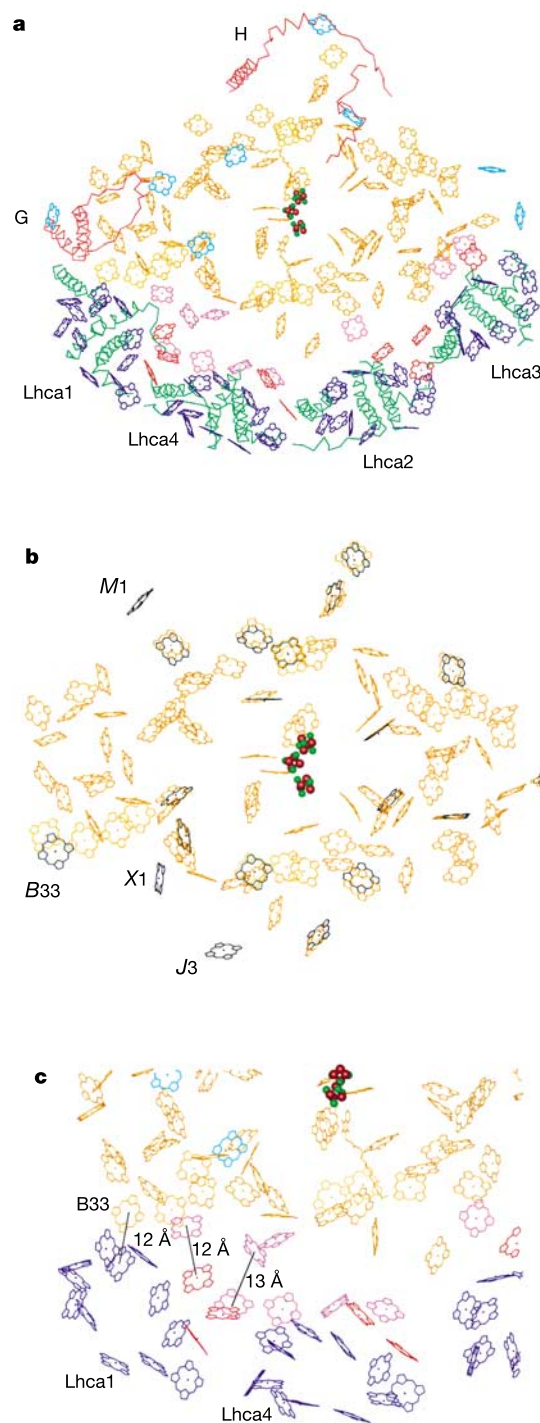


Figure 2 The arrangement of 167 chlorophyll molecules of plant PSI as seen from the stromal side. Yellow, plant reaction centre chlorophylls that are present in cyanobacteria; cyan, reaction centre chlorophylls unique for plants; blue, chlorophylls bound to the Lhca monomers; red, LHCI linker chlorophylls (see text); magenta, chlorophylls positioned in the cleft between LHCI and the reaction centre (designated as gap chlorophylls in the text). **a**, Chlorophyll arrangement including $\text{C}\alpha$ backbone of subunits that are exclusive to eukaryotic PSI. **b**, Conservation of cyanobacterial reaction centre chlorophylls in plant. The cyanobacterial chlorophylls that are missing or that have changed their position or orientation in plant reaction centre are in black. **c**, The contact region between the LHCI belt and the reaction centre in the vicinity of subunit G and Lhca1. Shortest $\text{Mg}^{2+}\text{--Mg}^{2+}$ distances between LHCI belt and core chlorophylls are indicated.

LHCI dimers? We find that helix D at the luminal carboxy-terminal domain of Lhca1 protrudes out of the main body of this protein and binds to both helix C and the luminal loop, connecting helices C and B, of Lhca4 (Fig. 3a). A similar, but apparently weaker, binding mode appears between Lhca2 and Lhca3, and also between the two dimers. The mode of binding depicted in Fig. 3a underlines the importance of the C-terminal domain of Lhca1 for dimer formation¹⁶, and suggests a similar role for the N terminus. Hence, the LHCI proteins are arranged in a series where helix D of one monomer is pointing towards helix C of the next. This maximizes the number of LHCI chlorophylls facing the core. It is noteworthy that the loops connecting the transmembrane helices are the least conserved between LHCI and trimer-forming LHCII^{23,24}. In spite of this dimerization mode, which does not involve interactions between transmembrane helices, the interpigment distances between the nearest chlorophylls of neighbouring monomers are still very short. This is due to the 'linker' chlorophylls, located between the monomers, which may have an important role in energy migration along the LHCI belt (Figs 2a and 3a). The possible physical interaction with chlorophylls at the interfaces between LHCI monomers suggests that the linker chlorophylls could also be important for dimer formation.

As expected from the amino acid sequences, the helix structure of LHCI (Fig. 4a–d) and LHCII²³ is quite similar. All four LHCI monomers share the LHCII general fold; that is, two long, tilted, intertwined transmembrane helices (A and B), a shorter one roughly perpendicular to the membrane (C), and a 10–12-amino-acid-long hydrophilic helix parallel to the luminal membrane (D). In terms of absorption characteristics, the main distinction between LHCI and LHCII is the longer wavelength absorption spectrum.

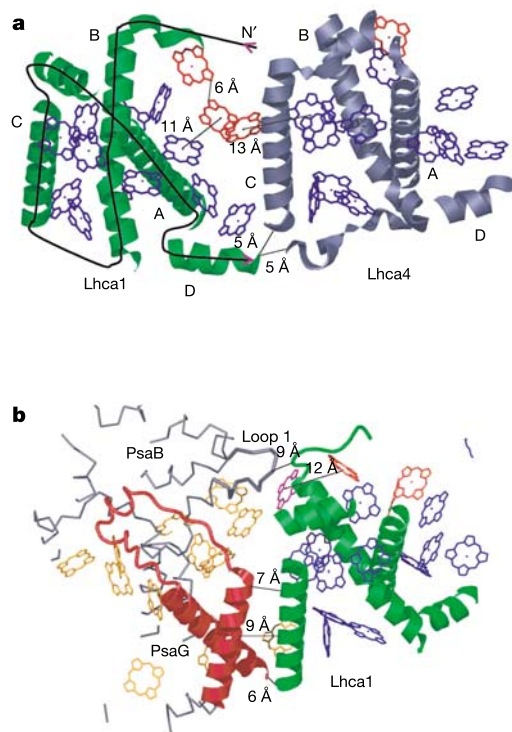


Figure 3 Dimer formation between Lhca1 and Lhca4 and tight binding of Lhca1 to the reaction centre. **a**, The close contact between helix D of Lhca1 (green) and both the luminal region of helix C and the luminal loop connecting helices B and C of Lhca4 (grey) is shown. The magenta solid line follows schematically the Lhca1 backbone from the putative N terminus (N') to the C terminus. Chlorophyll colour code is as in Fig. 2. The nomenclature for the four helices (A, B, C and D) is adopted from LHCII²³. **b**, Tight binding of Lhca1 to subunits B and G of the reaction centre.

This difference is mainly attributable to tighter interactions between pigments in LHCI; however, protein–pigment interactions may also contribute²⁵. The structure reveals that although the positions, number of chlorophylls and Mg^{2+} – Mg^{2+} distances in LHCI monomers are similar to those in LHCII, there are a few marked alterations particularly in chromophore orientations (Fig. 4c, d). Most noticeable are the sharp differences in the tilting of chlorophylls *b5* (or *a5*), *a7* and *a3* (see ref. 23 for nomenclature), as well as the modified location of chlorophyll *b2*, which in LHCI is positioned closer and parallel to a linker chlorophyll located between two monomers. All these chlorophylls face either the core or the neighbouring monomer. The most prominent distinction in chlorophyll arrangement between LHCI and LHCII is not to be found in the isolated monomers (Fig. 4b, d). Within each dimer, two linker chlorophylls are bound at the interface between the neighbouring monomers and one chlorophyll is located between the two dimers. In addition, another linker chlorophyll, which appears in all four LHCI proteins, is positioned close to chlorophyll *a4* and faces the reaction centre.

Binding of LHCI to the reaction centre is asymmetric; that is, much stronger on the G-pole than on the K-pole of the core (Fig. 1a). Lhca1 is strongly attached to the core through the helix bundle formed between its transmembrane helix C and the two tilted helices of PsaG, and owing to the close interaction of its stromal loop with the novel loop 1 of PsaB (Fig. 3b). The other LHCI proteins interact with the core mainly through small binding surfaces at their stromal-exposed regions (Fig. 1a). Lhca4 binds to PsaF, Lhca2 associates weakly with PsaJ, and Lhca3 interacts with an unassigned electron density that is probably attributable to the stromal loop of PsaK (on the luminal side Lhca3 binds weakly to PsaA). This observation fits in with recent work concerning alterations of LHCI composition with changing environmental conditions^{7,26}. We suggest that Lhca1, and to some extent the Lhca1–Lhca4 dimer, act as an anchor point for facilitating the binding of other LHCI monomers and dimers (and their isoforms) at varying

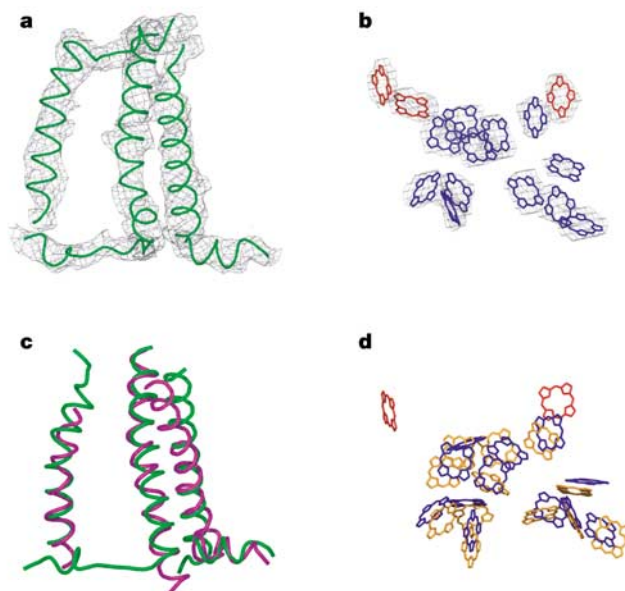


Figure 4 The structural model of Lhca monomers compared to LHCII. **a**, The experimental electron density map covering the $\text{C}\alpha$ backbone of Lhca4. **b**, Experimental electron density map covering the chlorophylls of Lhca4. Chlorophylls in blue have parallels in LHCII whereas the additional linker chlorophylls (see text) are in red. **c**, Superposition of Lhca2 (green) and LHCII (magenta) backbones. **d**, Superposition of the chlorophylls of Lhca2 (blue and red) and LHCII (yellow). Note that this Lhca contains only two linker chlorophylls.

stoichiometries, depending on environmental conditions. Our structure therefore suggests that an important function unique to each Lhca protein is the fine-tuned and tailor-made interactions that it has with the core and other Lhca monomers.

Chlorophyll arrangement and excitonic energy transfer

The four LHCI proteins are evenly spaced along the antenna belt, with virtually identical distances between dimers and dimer subunits (Fig. 1a). The belt is densely populated with chlorophylls. Within the membrane, the C α backbone of LHCI maintains a distance of more than 20 Å from most parts of the reaction centre. Owing to this gap, most LHCI chlorophylls are separated by more than 18 Å from the nearest reaction centre chlorophyll (Mg²⁺–Mg²⁺ distances). However, there are three contact regions where much shorter interpigment distances (10–15 Å) are observed, and these are proposed to play an important role in energy migration. These denser contact regions are located at the two poles of the LHCI belt near PsaK and PsaG (Fig. 2a, c), with an additional minor region located at the centre of the LHCI belt in the vicinity of PsaF. Core chlorophylls that are not present in cyanobacteria, or antenna chlorophylls that have no counterparts in LHCI, are involved in most of the chlorophyll pairs forming the closer interactions. Apparently, optimization of energy transfer between LHCI and the core arose through the evolution of a small number of strategically positioned chlorophyll-binding sites that we designate as ‘gap’ chlorophylls. Except for the described linker chlorophylls, which are intimately associated with LHCI, ten additional gap chlorophylls are not an intrinsic part of the LHCI belt or the reaction centre. They are bound to the periphery of either one or both of these moieties and are positioned within the cleft between them. The non-uniform distribution of close LHCI–core contacts among the three contact regions suggests the existence of distinct energy migration routes from the antenna to the core of PSI.

The chlorophyll arrangement described here, where most antenna chlorophylls are located 20–32 Å from core chlorophylls (Mg²⁺–Mg²⁺ distance) except at the contact regions, resembles the architecture of chlorophyll distribution in the reaction centre.

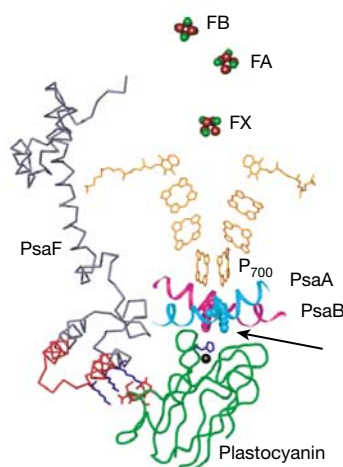


Figure 5 Electron transfer chain and plastocyanin binding. Residues 18–47 of plant PsaF, which include the 18 residues exclusive to eukaryotes and that form a helix–loop–helix domain, are coloured red. A cluster of four negatively charged residues (red) on the surface of plastocyanin (green) interacts with three lysines (blue) from the N terminus of PsaF, two of which are from the extra 18 residues. The two tryptophan residues (indicated by an arrow) crucial for the electron transfer from the reduced copper to the oxidized P₇₀₀, and that contribute to the hydrophobic interaction with plastocyanin, are shown as spherical atoms (cyan and magenta). The histidine residue that coordinates the copper atom in plastocyanin is in blue.

There, the innermost chlorophylls are all distanced 20–30 Å from the electron transfer chain, except for two linker chlorophylls that are positioned at greater proximity to it. This arrangement maximizes the number of chlorophylls that can deliver energy directly to the electron transfer chain.

Here we show that the LHCI belt is much more densely populated than previously estimated²⁷. With its 56 chlorophylls (not including gap chlorophylls), the ratio of chlorophyll/protein exceeds even that of the reaction centre or LHCII. Such a high pigment density, yielding a more intricate set of close pigment–pigment interactions (Fig. 2a), may be responsible for LHCI’s unique spectral characteristics. LHCI harbours most of the red-shifted chlorophylls of plant PSI²⁸. These are of considerable importance for light harvesting in dense vegetation systems, where ambient light is enriched in wavelengths above 690 nm (ref. 29). Energy from these chlorophylls migrates to neighbouring bulk chlorophylls through a slow, thermally activated transfer that reduces the rate of energy migration from the antenna to the core³⁰. The enhanced pigment–pigment

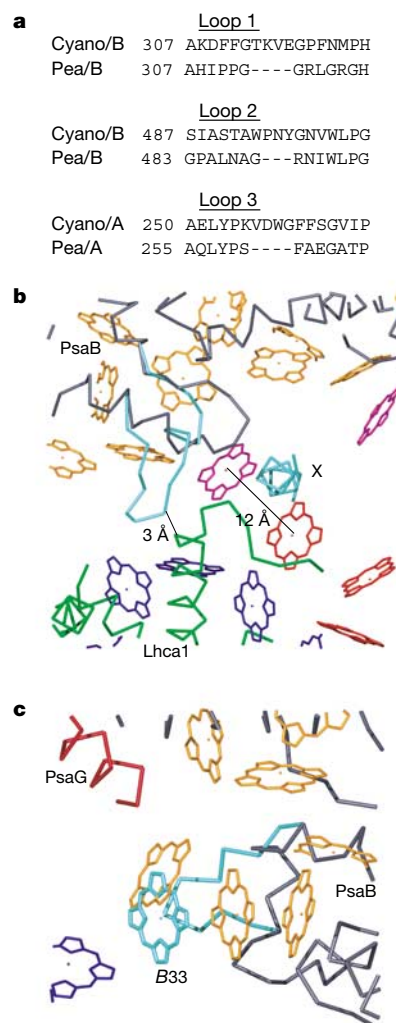


Figure 6 Loops altered during the evolution of plant PsaA and PsaB. The cyanobacterial domains modified or eliminated in plants are in cyan; plant C α backbone is in grey; and the colour code of the various plant chlorophylls is as in Fig. 2. **a**, Amino acid sequence alignment of the regions that accommodate the three altered loops. **b**, Loop 1. This stromal loop was altered and subunit X was eliminated in the plant reaction centre. The altered loop coordinates a gap chlorophyll positioned close to Lhca1 chlorophyll (distance indicated). **c**, Loop 2. Modification of loop 2 results in a large change in the position of chlorophyll B33 that in *S. elongatus* (cyan) forms part of a trimer of parallel chlorophylls that was proposed to absorb light at longer wavelengths.

interaction in LHCI may decrease the time required for the slow 'uphill' push from low-energy chlorophylls by providing many possible acceptors in their vicinity.

Plastocyanin binding and electron transfer

Plastocyanin is the only electron donor of plant PSI, and PsfF is known to provide part of the plastocyanin-binding site^{20,31}. Figure 5 depicts a model of plastocyanin- and plant PSI-binding, based on our model of PSI and the available structure of plastocyanin (Protein Data Bank code 1ag6)³². In this model, plastocyanin's positioning in the putative binding site was guided by hydrophobic interactions between a region of plastocyanin, proximal to the copper ion, with the luminal loops containing Trp 625 in PsfB and Trp 658 in PsfA (Fig. 5 and ref. 33). Superfluous degrees of freedom have been resolved by bringing into contact a cluster of negatively charged conserved residues (Asp 42, Glu 43, Asp 44 Glu 45) of plastocyanin with the positively charged N-terminal domain of PsfF, which contains a few lysine residues that are not present in cyanobacteria^{20,34}. This arrangement brings the copper atom in plastocyanin to coincide with the pseudo two-fold axis of symmetry of the electron transfer path from the special chlorophyll pair (P₇₀₀) to the Fe₄-S₄ cluster Fx.

Electron transfer from plastocyanin to PSI is two orders of magnitude faster in plants compared with cyanobacteria²⁰. This is probably due to more efficient plastocyanin binding in plants, mediated by the extra 18 amino acid residues in the N terminus of plant PsfF (Fig. 1b). This extra N-terminal domain forms an amphipathic helix-loop-helix motif on the luminal side of the thylakoid membrane (Fig. 5). Sequence alignment and the good match between our electron density map and the cyanobacterial model outside this region allowed us to place the known sequence of the extra 18 residues within this luminal helix-loop-helix. One of these helices contains two lysine residues—Lys 30 and to a greater extent Lys 23—that are responsible for the much stronger electrostatic interaction between plastocyanin and PSI³⁵.

Evolutionary forces shaping plant PSI

Sequence alignment of the core main subunits A and B between plants and *S. elongatus* revealed only three sites where amino acids were lost in plants (Fig. 6a). They all form parts of solvent-exposed loops denoted loop 1, loop 2 and loop 3. As depicted in Fig. 6b, closer association of Lhca1 with the core probably arose through modification of the protruding cyanobacterial loop 1 and elimination of PsfX. The resulting smaller loop binds Lhca1 and coordinates a gap chlorophyll that might have a role in energy migration from this antenna protein to the core. Similarly, modification in the protruding cyanobacterial loop 3 enabled closer association of Lhca3 with the core (not shown). Alteration of loop 2 (Fig. 6c) results in a substantial change in the position of chlorophyll B33 that, together with B31 and B32, forms in *S. elongatus* a strongly coupled trimer to which absorption at the end of its range in the red-light wavelength is attributed¹⁰. This modification is therefore probably manifested in the different absorption spectra of the plant reaction centre. This is the only instance where the position of cyanobacterial chlorophyll changed significantly in the plant core: note that it takes place at one of the contact regions.

Whereas plant PSI is purified as a monomer its cyanobacterial counterpart assembles *in vivo* into trimers¹. By superimposing three plant PSI monomers on top of the cyanobacterial trimer we are able to conclude unequivocally that plant PSI cannot form similar trimers because its PsfH, not present in cyanobacteria, completely hinders the formation of contacts among the monomers. As PsfH probably forms part of the docking site for LHCI it implies that trimerization was lost in plants to facilitate re-allocation of phosphorylated LHCI to PSI under light conditions favouring PSII excitation. Furthermore, we find that the C terminus of PsfL, which protrudes in the cyanobacterial reaction centre and facilitates trimer

formation, is lost in plants. Again, as in loop 1 and loop 3, we suggest that association with a membrane antenna protein, in this case LHCI, forced the elimination of a protruding domain.

Recently it was discovered that under certain stress conditions, not uncommon to a marine habitat, the cyanobacterial PSI is also augmented by a membrane antenna protein called CP43' (unrelated to LHCI monomers but related to CP43 from PSII). Eighteen identical copies of CP43' form a circle around the periphery of the PSI trimer, which keeps an almost constant distance from the core (on its F/J side) except for three closer interaction regions per monomer^{36–38}. As this circle is energetically efficiently coupled to the reaction centre³⁹, we hypothesize that the cyanobacterial reaction centre is optimized not only for a 'stand alone' set up but also for trapping energy delivered from a peripheral membrane antenna. This might explain the paucity of alterations that plant reaction centre chlorophylls had to undergo.

The major differences in the overall architecture of the LHCI and CP43' belts may reflect the need of plants to adjust to an environment with a different light regime and where light intensity and quality are subject to much greater dynamic variations^{7,29}. In the LHCI belt, chlorophyll density is almost double that of the CP43' belt. This higher density gave rise to the red-shifted chlorophylls that broadened the plant PSI absorption spectra. It also enabled placing the belt of chlorophylls much closer to the core (interpigment distances between core and belt in the three contact regions are 20–25 Å in PSI–CP43' compared with 10–15 Å in LHCI–PSI). This proximity, made possible by eliminating some cyanobacterial protruding loops, might be essential in compensating for the red chlorophylls that significantly reduce the rate of energy migration to the core. LHCI is composed of four different building blocks that change their composition in response to varying environmental conditions. This flexibility in LHCI stoichiometry is facilitated by the strong binding of Lhca1 to PsfG, the looser association of the other Lhca monomers to the core, and the relatively weak interaction between the Lhca proteins.

The structure of plant PSI presented here reveals the spatial distribution of all of its transmembrane helices and the vast majority of the cofactors involved in energy and electron transfer. It provides a framework for further investigation of the mechanisms regulating the composition and activity of PSI's numerous subunits under changing environmental conditions. □

Methods

PSI was isolated as a monomer from fresh pea (*Pisum sativum* var. *alaska*) leaves and crystallized as described elsewhere^{11,40} in space group P2₁, with unit-cell parameters $a = 181.90 \text{ Å}$, $b = 190.24 \text{ Å}$, $c = 219.66 \text{ Å}$, $\beta = 90.48^\circ$, exhibiting diffraction to 4 Å resolution at a synchrotron source. Phases were determined using several heavy-atom derivatives and the anomalous signal of the intrinsic Fe–S clusters. Figure of merit converged to 0.43 before and 0.71 after density modification. The asymmetric unit contains two PSI copies related by non-crystallographic symmetry that cannot occur *in vivo* within the membrane. The final electron density map allowed tracing of all transmembrane helices (modelled by polyalanine), and some additional structural elements, as well as determination of the position and tilting of the vast majority of chlorophylls (modelled by porphyrins). The final model was refined as a rigid body yielding $R = 0.41$ and $R_{\text{free}} = 0.42$. Supplementary Table 1 provides detailed information about the X-ray data collection and phasing. The conserved parts of plant and cyanobacterial reaction centres were handled as described in the Supplementary Methods.

Received 21 August; accepted 28 October 2003; doi:10.1038/nature02200.

1. Chittis, P. R. Photosystem I: Function and physiology. *Annu. Rev. Plant Physiol. Plant. Mol. Biol.* **52**, 593–626 (2001).
2. Barber, J. Photosystem II: a multisubunit membrane protein that oxidises water. *Curr. Opin. Struct. Biol.* **12**, 523–530 (2002).
3. Junge, W. ATP synthase and other motor proteins. *Proc. Natl Acad. Sci. USA* **96**, 4735–4737 (1999).
4. Cramer, W. A. *et al.* Some new structural aspects and old controversies concerning the cytochrome b(6)f complex of oxygenic photosynthesis. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **47**, 477–508 (1996).
5. Herrmann, R. G. Biogenesis and evolution of photosynthetic (thylakoid) membranes. *Biosci. Rep.* **19**, 355–365 (1999).
6. Trissl, H.-W. & Wilhelm, C. Why do thylakoid membranes from higher plants form grana stacks? *Trends Biochem. Sci.* **18**, 415–419 (1993).
7. Bailey, S., Walters, R. G., Jansson, S. & Horton, P. Acclimation of *Arabidopsis thaliana* to the light environment: the existence of separate low light and high light responses. *Planta* **213**, 794–801 (2001).

8. Durnford, D. G. *et al.* A phylogenetic assessment of the eukaryotic light-harvesting antenna proteins, with implications for plastid evolution. *J. Mol. Evol.* **48**, 59–68 (1999).
9. Croce, R., Morosinotto, T., Castelletti, S., Breton, J. & Bassi, R. The Lhca antenna complexes of higher plants photosystem I. *Biochim. Biophys. Acta* **1556**, 29–40 (2002).
10. Jordan, P. *et al.* Three-dimensional structure of cyanobacterial photosystem I at 2.5 Å resolution. *Nature* **411**, 909–917 (2001).
11. Ben-Shem, A., Nelson, N. & Frolow, F. Crystallization and initial X-ray diffraction studies of higher plant photosystem I. *Acta Crystallogr. D* **59**, 1824–1827 (2003).
12. Scheller, H. V., Jensen, P. E., Haldrup, A., Lunde, C. & Knoetzel, J. Role of subunits in eukaryotic Photosystem I. *Biochim. Biophys. Acta* **1507**, 41–60 (2001).
13. Jansson, S., Andersen, B. & Scheller, H. V. Nearest-neighbor analysis of higher-plant photosystem I holocomplex. *Plant Physiol.* **112**, 409–420 (1996).
14. Kargul, J., Nield, J. & Barber, J. Three-dimensional reconstruction of a light-harvesting complex I-photosystem I (LHCI-PSI) supercomplex from the green alga *Chlamydomonas reinhardtii*. Insights into light harvesting for PSI. *J. Biol. Chem.* **278**, 16135–16141 (2003).
15. Ihalaenen, J. A. *et al.* Pigment organization and energy transfer dynamics in isolated photosystem I (PSI) complexes from *Arabidopsis thaliana* depleted of the PSI-G, PSI-K, PSI-L, or PSI-N subunit. *Biophys. J.* **83**, 2190–2201 (2002).
16. Schmid, V. H., Paulsen, H. & Rupprecht, J. Identification of N- and C-terminal amino acids of Lhca1 and Lhca4 required for formation of the heterodimeric peripheral photosystem I antenna LHCI-730. *Biochemistry* **41**, 9126–9131 (2002).
17. Ganeteg, U., Strand, A., Gustafsson, P. & Jansson, S. The properties of the chlorophyll a/b-binding proteins Lhca2 and Lhca3 studied *in vivo* using antisense inhibition. *Plant Physiol.* **127**, 150–158 (2001).
18. Knoetzel, J., Mant, A., Haldrup, A., Jensen, P. E. & Scheller, H. V. PSI-O, a new 10-kDa subunit of eukaryotic photosystem I. *FEBS Lett.* **510**, 145–148 (2002).
19. Lunde, C. P., Jensen, P. E., Haldrup, A., Knoetzel, J. & Scheller, H. V. The PSI-H subunit of photosystem I is essential for state transitions in plant photosynthesis. *Nature* **408**, 613–615 (2000).
20. Hippler, M. *et al.* The plastocyanin binding domain of photosystem I. *EMBO J.* **15**, 6374–6384 (1996).
21. Raymond, J., Zhaxybayeva, O., Gogarten, J. P., Gerdes, S. Y. & Blankenship, R. E. Whole-genome analysis of photosynthetic prokaryotes. *Science* **298**, 1616–1620 (2002).
22. Xiong, J. & Bauer, C. E. Complex evolution of photosynthesis. *Annu. Rev. Plant Biol.* **53**, 503–521 (2002).
23. Kuhlbrandt, W., Wang, D. N. & Fujiyoshi, Y. Atomic model of plant light-harvesting complex by electron crystallography. *Nature* **367**, 614–621 (1994).
24. Jansson, S. A. Guide to the Lhc genes and their relatives in *Arabidopsis*. *Trends Plant Sci.* **4**, 236–240 (1999).
25. Morosinotto, T., Castelletti, S., Breton, J., Bassi, R. & Croce, R. Mutation analysis of Lhca1 antenna complex. Low energy absorption forms originate from pigment–pigment interactions. *J. Biol. Chem.* **277**, 36253–36261 (2002).
26. Moseley, J. L. *et al.* Adaptation to Fe-deficiency requires remodeling of the photosynthetic apparatus. *EMBO J.* **21**, 6709–6720 (2002).
27. Schmid, V. H. *et al.* Pigment binding of photosystem I light-harvesting proteins. *J. Biol. Chem.* **277**, 37307–37314 (2002).
28. Croce, R., Zucchelli, G., Garlaschi, F. M. & Jennings, R. C. A thermal broadening study of the antenna chlorophylls in PSI-200, LHCI, and PSI core. *Biochemistry* **37**, 17355–17360 (1998).
29. Rivadossi, A., Zucchelli, G., Garlaschi, F. M. & Jennings, R. C. The importance of PSI chlorophyll red forms in light-harvesting by leaves. *Photosynth. Res.* **60**, 209–215 (1999).
30. Jennings, R. C., Zucchelli, G., Croce, R. & Garlaschi, F. M. The photochemical trapping rate from red spectral states in PSI-LHCI is determined by thermal activation of energy transfer to bulk chlorophylls. *Biochim. Biophys. Acta* **1557**, 91–98 (2003).
31. Bengis, C. & Nelson, N. Subunit structure of chloroplast photosystem I reaction center. *J. Biol. Chem.* **252**, 4564–4569 (1977).
32. Xue, Y., Okvist, M., Hansson, O. & Young, S. Crystal structure of spinach plastocyanin at 1.7 Å resolution. *Protein Sci.* **7**, 2099–2105 (1998).
33. Sommer, F., Drepper, F. & Hippler, M. The luminal helix I of PsaB is essential for recognition of plastocyanin or cytochrome c6 and fast electron transfer to photosystem I in *Chlamydomonas reinhardtii*. *J. Biol. Chem.* **277**, 6573–6581 (2002).
34. Chitnis, P. R., Purvis, D. & Nelson, N. Molecular cloning and targeted mutagenesis of the gene *psaF* encoding subunit III of photosystem I from the cyanobacterium *Synechocystis* sp. PCC 6803. *J. Biol. Chem.* **266**, 20146–20151 (1991).
35. Hippler, M., Drepper, F., Haehnel, W. & Roichaix, J. D. The N-terminal domain of PsaF: precise recognition site for binding and fast electron transfer from cytochrome c6 and plastocyanin to photosystem I of *Chlamydomonas reinhardtii*. *Proc. Natl Acad. Sci. USA* **95**, 7339–7344 (1998).
36. Bibby, T. S., Nield, J. & Barber, J. Iron deficiency induces the formation of an antenna ring around trimeric photosystem I in cyanobacteria. *Nature* **412**, 743–745 (2001).
37. Boekema, E. J. *et al.* A giant chlorophyll–protein complex induced by iron deficiency in cyanobacteria. *Nature* **412**, 745–748 (2001).
38. Nield, J., Morris, E. P., Bibby, T. S. & Barber, J. Structural analysis of the photosystem I supercomplex of cyanobacteria induced by iron deficiency. *Biochemistry* **42**, 3180–3188 (2003).
39. Melkozernov, A. N., Lin, S., Bibby, T. S., Barber, J. & Blankenship, R. E. Time-resolved absorption and emission show that the CP43' antenna ring of iron-stressed *Synechocystis* sp. PCC6803 is efficiently coupled to the photosystem I reaction center core. *Biochemistry* **42**, 3893–3903 (2003).
40. Nelson, N. & Ben-Shem, A. Photosystem I reaction center: Past and future. *Photosynth. Res.* **73**, 193–206 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We acknowledge the ESRF for synchrotron beam time, and staff scientists of the ID14 stations cluster for their assistance. We thank W. Kuhlbrandt for LHCII coordinates. A.B. is a recipient of a Charles Clore Foundation Ph.D. student scholarship. This work was supported by a grant from The Israel Science Foundation to N.N. and F.F.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.N. (nelson@post.tau.ac.il). Atomic coordinates are deposited in the Protein Data Bank under accession number 1qzv.

The essential signature of a massive starburst in a distant quasar

P. Solomon¹, P. Vanden Bout², C. Carilli³ & M. Guelin⁴

¹Department of Physics and Astronomy, SUNY at Stony Brook, Stony Brook, New York 11794, USA

²National Radio Astronomy Observatory, 520 Edgemont Road, Charlottesville, Virginia 22903, USA

³National Radio Astronomy Observatory, PO Box 0, Socorro, New Mexico 87801, USA

⁴Institut de Radioastronomie Millimétrique, 300 rue de la Piscine, Domaine Universitaire, 38406 Saint Martin d'Hères, France

Observations of carbon monoxide emission in high-redshift ($z > 2$) galaxies indicate the presence of large amounts of molecular gas. Many of these galaxies contain an active galactic nucleus powered by accretion of gas onto a supermassive black hole, and a key question is whether their extremely high infrared luminosities result from the active galactic nucleus, from bursts of massive star formation (associated with the molecular gas), or both. In the Milky Way, high-mass stars form in the dense cores of interstellar molecular clouds, where gas densities are $n(\text{H}_2) > 10^5 \text{ cm}^{-3}$ (refs 1, 2). Recent surveys show that virtually all galactic sites of high-mass star formation have similarly high densities³. The bulk of the cloud material traced by CO observations, however, is at a much lower density. For galaxies in the

local Universe, the HCN molecule is an effective tracer of high-density molecular gas⁴. Here we report observations of HCN emission from the infrared-luminous 'Cloverleaf' quasar (at a redshift $z = 2.5579$). The HCN line luminosity indicates the presence of 10 billion solar masses of very dense gas, an essential feature of an immense starburst, which contributes, together with the active galactic nucleus it harbours, to its high infrared luminosity.

The Cloverleaf is one of about 25 galaxies observed in the epoch of galaxy formation ($z > 2$) in carbon monoxide (CO) emission. The Cloverleaf (also known as H1413+117) has its optical and CO emission imaged into four components by a gravitational lens. The first detection of the Cloverleaf in molecular line emission was that of CO($J = 3-2$) (ref. 5). It has now been observed⁶ in four transitions of CO, with a weighted mean redshift for the CO emission of $z = 2.5579$, and also in the atomic carbon fine-structure lines^{6,7}. A recently constructed source-lens model (see ref. 8 for a discussion of this and other Cloverleaf lens models), based on high-resolution images of CO($J = 7-6$) emission, shows that the Cloverleaf contains an extensive molecular disk with a diameter of 1.6 kpc.

CO emission, in this source and in general, is the best tracer of molecular clouds, the potential fuel for star formation and a major component of interstellar matter. But the low density required for excitation of CO, $n(\text{H}_2) > 300 \text{ cm}^{-3}$, means that it is not a specific tracer of molecular cloud cores or star formation rates. In contrast, HCN is a specific tracer of cloud cores. The first observed correlation⁴ between HCN($J = 1-0$) luminosity and infrared luminosity (L_{IR}) for galaxies has recently been expanded to cover many more

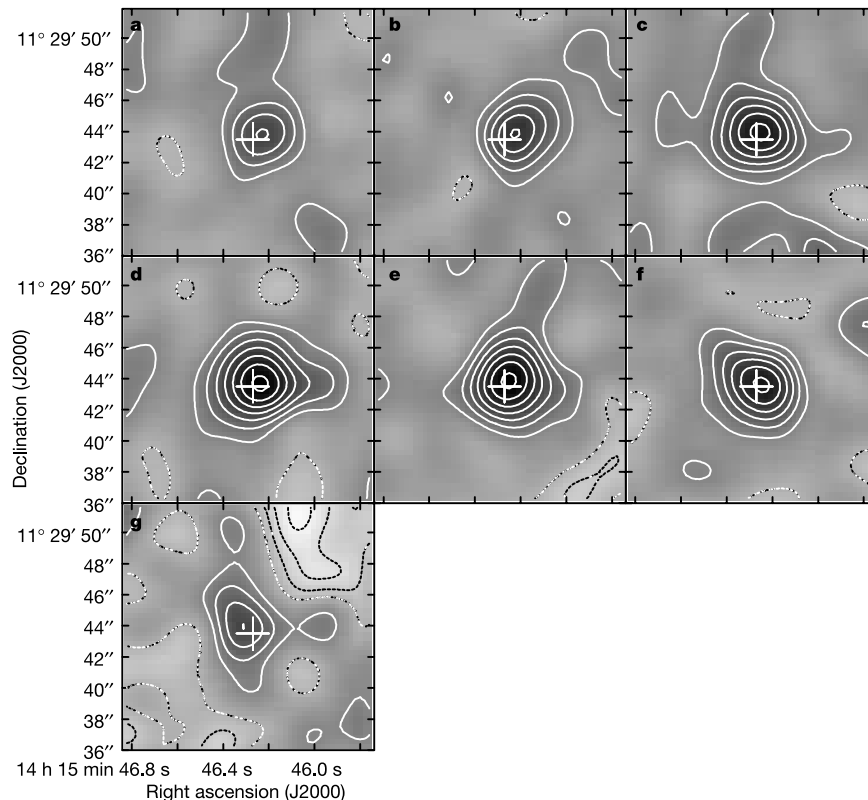


Figure 1 HCN($J = 1-0$) channel maps from the Cloverleaf observations obtained with the NRAO Very Large Array. Contour levels are $-3, -2, -1, 1, 2, 3, 4, 5, 6, 7$ and $8 \times 0.07 \text{ mJy per beam}$ (equivalent to 1.2 times the r.m.s. noise, except for channel 7 which has an r.m.s. noise of 0.09 mJy). Dash-dot contours are negative; solid contours are positive. The continuum has not been subtracted. The naturally weighted beam has a full-width at half-maximum of 4 arcsec. **a–g**, The channel width is 6.25 MHz, equivalent

to 75 km s^{-1} ; the channels are centred at $+231, +154, \dots, -231 \text{ km s}^{-1}$ with respect to an observed frequency of 24.9149 GHz . The coordinates of the position marked by a cross are right ascension (RA) $14 \text{ h } 15 \text{ min } 46.27 \text{ s}$, declination (dec.) $+11^\circ 29' 43.5''$; J2000. The observations were made in the D-configuration of the VLA on 10, 16 and 31 March 2003, with total integration time of 12 h.

sources⁹. The HCN–IR correlation for galaxies, which extends over three orders of magnitude in L_{IR} , is linear. Unlike $L_{\text{IR}}/L'_{\text{CO}}$, the ratio of $L_{\text{IR}}/L'_{\text{HCN}}$ does not increase with L_{IR} , indicating that the star formation rate per unit dense molecular gas is the same for ultraluminous galaxies and normal spiral galaxies. (Here L' indicates line luminosity; see below.) Observation of a high HCN line luminosity would strongly indicate the presence of high-mass star formation in the Cloverleaf, and to that end we have used the NRAO Very Large Array (VLA) to search for HCN($J = 1-0$) emission.

The VLA image (Fig. 1) of line plus continuum for seven spectral channels shows the line clearly at the position of the quasar in the four centre channels. The spectrum of the HCN emission (Fig. 2) has a similar profile to that of the CO($J = 7-6$) emission¹⁰. The HCN emission is clearly detected with a peak line flux density of six times the r.m.s. noise, $S_{\text{peak}} = 0.24 \pm 0.04$ mJy.

Table 1 summarizes the observed and derived (intrinsic) quantities, including the ratios $L_{\text{IR}}/L'_{\text{HCN}}$ and $L'_{\text{HCN}}/L'_{\text{CO}}$. The line luminosities L' are a measure of the surface brightness times area in brightness temperature units¹¹. Two lines with the same spatial extent, velocity profile, and brightness temperature T_b will have the same L' regardless of transition or line frequency. All the luminosities in Table 1 are corrected for lens magnification, using the results of a lens model⁸ based directly on a CO($J = 7-6$) image. This model leads to a derived CO disk diameter of 1.6 kpc, similar to that of the CO emitting regions in nearby starburst ultraluminous galaxies¹². It should be noted that the diameter of the dust emission region, responsible for producing the infrared (IR) luminosity, has been determined¹³ to be 1.5 kpc from a comparison of models of radiation transfer in the dust with the observed spectral energy distribution. Thus, both the molecular and IR emission originate in extended regions of the same size. In this respect, the Cloverleaf is similar to PSS J2322 + 1944, where the molecular gas in a ring surrounding a quasar has been directly imaged using gravitational lensing¹⁴.

The Cloverleaf HCN($1-0$) line luminosity is larger by a factor of 100 than that of normal spiral galaxies. The only galaxies with HCN line luminosities comparable to that of the Cloverleaf are ultraluminous IR galaxies. The intrinsic HCN line luminosity of the Cloverleaf is a factor of about two higher than that of the low-redshift ultraluminous IR galaxy Mrk231, the most luminous object in the local Universe, and a factor of three higher than that of Arp220, the prototype ultraluminous IR galaxy. To put this in perspective, the Cloverleaf intrinsic HCN line luminosity is ten times greater than the CO line luminosity of the Milky Way. A large HCN line luminosity is a clear indicator of a large mass of dense star-forming molecular gas. Assuming that the HCN emission originates in dense gravitationally bound regions of average density

$n(\text{H}_2) = 3 \times 10^4 \text{ cm}^{-3}$ and intrinsic brightness temperature $T_b = 50 \text{ K}$, the mass of dense gas is about $7 \times L'_{\text{HCN}}$, or 2×10^{10} solar masses. (The gas mass scales as $n^{1/2}/T_b$ for gravitationally bound clouds.)

In star formation models, the star formation rate is proportional to the IR luminosity of dust heated by embedded recently formed high-mass stars. The ratio $L_{\text{FIR}}/L'_{\text{HCN}}$ is a measure of the star formation rate per unit mass of dense gas⁴ (L_{FIR} , far-IR luminosity). L'_{CO} is an indicator of the total molecular gas, because the CO transitions are so easily excited. The ratio $L'_{\text{HCN}}/L'_{\text{CO}}$ is a measure of the fraction of all molecular gas that is in dense star-forming cloud cores. This fraction is much higher in ultraluminous IR galaxies than in normal spirals. For the Cloverleaf, $L_{\text{IR}}/L'_{\text{HCN}}$ is about a factor of 5 times higher than typical ultraluminous IR galaxies, and $L'_{\text{HCN}}/L'_{\text{CO}}$ is half that of Arp220 and a factor of 3 less than that of Mrk231. In making this comparison, we have assumed that the brightness temperature of the CO($J = 1-0$) line is 1.1 times the $J = 3-2$ line⁷. (One possibility that we have not considered is that the magnification for the molecular emission and the IR emission are different, which could distort the intrinsic ratios. The similar sizes deduced for the molecular disk and the dust emission is evidence that the magnification is the same for both.)

The molecular and IR luminosities for the Cloverleaf show that star formation in the large mass of dense molecular gas indicated by the HCN luminosity could account for a substantial fraction, but not all, of the IR luminosity from this quasar. Using Arp220 as a standard for the luminosity ratio $L_{\text{FIR}}/L'_{\text{HCN}}$, star formation in the dense molecular gas could account for 5×10^{12} times the solar luminosity or about 20% of the total intrinsic IR luminosity. Using normal spirals as a standard for the ratio^{4,9} $L_{\text{IR}}/L'_{\text{HCN}}$ would lead to a lower limit of 10%, while using the highest ratio for an ultraluminous IR galaxy gives an upper limit of 40%. A recent model⁷ of the IR spectral energy distribution of the Cloverleaf has two distinct components: one with a warm dust temperature $T_d = 115 \text{ K}$ responsible for the mid-infrared, and a second much more massive

Table 1 Observed and derived (intrinsic) quantities

Observed quantities: HCN($1-0$)	H1413+117	Mrk231	Arp220
z (HCN)*	2.5569 ± 0.0006		
S (HCN, peak)† (mJy)	0.24 ± 0.04		
I (HCN) = $S(\Delta v)$; (Jy km s ⁻¹)	0.069 ± 0.012		
Derived‡ intrinsic quantities			
L' (HCN, $J = 1-0$); (K km s ⁻¹ pc ²)	$(3.2 \pm 0.5) \times 10^9$	2×10^9	1.1×10^9
$L_{\text{IR}}/L'_{\text{HCN}}$; ($L_{\text{IR}}/\text{K km s}^{-1} \text{ pc}^2$)	$7,700 \pm 1,300$	1,600	1,500
$L_{\text{FIR}}/L'_{\text{HCN}}$; ($L_{\text{FIR}}/\text{K km s}^{-1} \text{ pc}^2$)	$1,700 \pm 300$	1,100	1,300
$L_{\text{HCN}}/L'_{\text{CO}}§$	0.07 ± 0.01	0.24	0.12

*Taking line centre halfway between two peak channels; rest frequency of 88.632 GHz.

†A tentative detection of the HCN($J = 4-3$) line has been reported (ref. 6) with a peak line flux density of $S = 5.5 \pm 1.1$ mJy, based on observations made with the IRAM 30-m telescope. Our own IRAM Interferometer observations of HCN($J = 4-3$) show no clear indication of a line. The interferometer data yields an upper limit of 3 mJy ($3 \times \text{r.m.s. noise}$) for the S(HCN 4-3) peak.

‡Throughout this Letter we have used a luminosity distance of 21 Gpc (flat universe with $H_0 = 70$, $\Omega_m = 0.3$ and $\Omega_\Lambda = 0.7$) and a lens magnification of 11 (ref. 8). Errors quoted for the derived quantities reflect the error in the precision of the integrated flux (I_{HCN}). There is also a possible systematic error from subtraction of the continuum with an uncertainty ($2 \times \text{r.m.s.}$) of 0.06 mJy. This would affect the accuracy of the integrated quantities by 25% (see Figs 1 and 2).

§Total infrared (IR; 10–2,000 μm), far-infrared (FIR; 42.5–122.5 μm) and CO($J = 1-0$) luminosities for the Cloverleaf (rest-frame wavelengths) from data in ref. 7; the IR/FIR and HCN luminosities in Mrk231 and Arp220 are taken from refs 9 and 4, respectively. L , luminosity of the Sun.

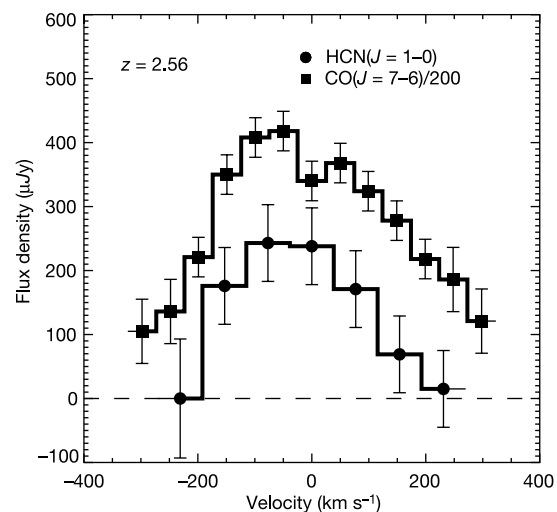


Figure 2 Spectrum of HCN($J = 1-0$) emission observed in the Cloverleaf together with the observed⁹ spectrum of CO($J = 7-6$), scaled down by a factor of 200. Zero velocity corresponds to a redshift $z = 2.5575$ at an observing frequency of 24.9149 GHz for HCN. The general similarity of the two line profiles, including the negative-velocity excess emission seen most clearly in the CO line, suggests common kinematics for the emission regions of the HCN line and this high-excitation CO line. A continuum of 0.26 mJy has been subtracted. The continuum was determined by an observation at 22.5 GHz and extrapolated to 24 GHz using the spectral index of -1.35 measured between 8 and 22 GHz. The result is $S_{\text{cont}}(24 \text{ GHz}) = 0.26 \pm 0.03$ mJy, where the error includes both the errors in the measurement at 22.5 GHz and the spectral index.

component with $T_d = 50$ K that produces the far-infrared. The far-infrared luminosity, 22% of the total, may correspond to the luminosity generated by star formation deduced above. The ratio $L_{\text{FIR}}/L_{\text{HCN}} = 1,700$ is then comparable to that of ultraluminous IR galaxies.

Adopting the Arp220 comparison for the expected ratio $L_{\text{FIR}}/L_{\text{HCN}}$ yields a star formation rate of 10^3 solar masses per year for the Cloverleaf, 300 times the rate in the Milky Way and 30 times greater than that of optical-ultraviolet starburst galaxies. The dense gas depletion time or starburst lifetime is about 10^7 years. Thus, in addition to the quasar, the Cloverleaf galaxy appears to have a huge starburst, more luminous than any optical starburst and comparable to that in ultraluminous IR galaxies. $\square$

Received 9 June; accepted 10 October 2003; doi:10.1038/nature02149.

1. Penzias, A. A., Solomon, P. M., Wilson, R. W. & Jefferts, K. B. Interstellar carbon monosulfide. *Astrophys. J.* **168**, L53–L58 (1971).
2. Solomon, P. M., Jefferts, K. B., Penzias, A. A. & Wilson, R. W. Interstellar methyl cyanide. *Astrophys. J.* **168**, L107–L110 (1971).
3. Plume, R., Jaffe, D. T., Evans, N. J. II, Martin-Pintado, J. & Gomez-Gonzalez, J. Dense gas and star formation: Characteristics of cloud cores associated with water masers. *Astrophys. J.* **476**, 730–749 (1997).
4. Solomon, P. M., Downes, D. & Radford, S. J. E. Dense molecular gas and starbursts in ultra-luminous galaxies. *Astrophys. J.* **387**, L55–L59 (1992).
5. Barvainis, R., Tacconi, L., Antonucci, R., Alloin, D. & Coleman, P. Extremely strong carbon-monoxide emission from the Cloverleaf quasar at a redshift of 2.5. *Nature* **371**, 586–588 (1994).
6. Barvainis, R., Maloney, P., Antonucci, R. & Alloin, D. Multiple CO transitions, C I, and HCN from the Cloverleaf quasar. *Astrophys. J.* **484**, 695–701 (1997).
7. Weiss, A., Henkel, C., Downes, D. & Walter, F. Atomic carbon in the Cloverleaf quasar at redshift of 2.5. *Astron. Astrophys.* (submitted).
8. Venturini, S. & Solomon, P. M. The molecular disk in the Cloverleaf quasar. *Astrophys. J.* **590**, 740–745 (2003).
9. Gao, Y. & Solomon, P. M. Star formation rate and dense molecular gas in galaxies. *Astrophys. J.* (submitted). Preprint at (<http://arXiv.org/astro-ph/0310339>) (2003).
10. Kneib, J.-P. *et al.* Modelling the Cloverleaf: Contribution of a galaxy cluster at $z \sim 1.7$. *Astron. Astrophys.* **329**, 827–839 (1998).
11. Solomon, P. M., Downes, D., Radford, S. J. E. & Barrett, J. W. The molecular interstellar medium in ultra-luminous infrared galaxies. *Astrophys. J.* **478**, 144–161 (1997).
12. Downes, D. & Solomon, P. M. Rotating nuclear rings and extreme starbursts in ultra-luminous galaxies. *Astrophys. J.* **507**, 615–654 (1998).
13. Granato, G. L., Danese, L. & Franceschini, A. Dust-enshrouded AGN models for hyperluminous, high-redshift infrared galaxies. *Astrophys. J.* **460**, L11–L14 (1996).
14. Carilli, C. *et al.* A molecular Einstein ring: Imaging a starburst disk surrounding a quasi-stellar object. *Science* **300**, 773–775 (2003).

Acknowledgements We thank D. Downes for discussions. The National Radio Astronomy Observatory is operated by Associated Universities, Inc., under cooperative agreement with the National Science Foundation. P.V.B. thanks Columbia University and the Institut d'Astrophysique de Paris for hospitality during this research.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.M.S. (psolomon@sbastk.ess.sunysb.edu).

Stationary pulses of light in an atomic medium

M. Bajcsy^{1,2}, A. S. Zibrov^{1,3,4} & M. D. Lukin¹

¹Physics Department, ²Division of Engineering and Applied Sciences, Harvard University, ³Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts 02138, USA

⁴Lebedev Institute of Physics, Moscow, 117924 Russia

Physical processes that could facilitate coherent control of light propagation are under active exploration^{1–5}. In addition to their fundamental interest, these efforts are stimulated by practical possibilities, such as the development of a quantum memory for photonic states^{6–8}. Controlled localization and storage of

photonic pulses may also allow novel approaches to manipulating of light via enhanced nonlinear optical processes⁹. Recently, electromagnetically induced transparency¹⁰ was used to reduce the group velocity of propagating light pulses^{11,12} and to reversibly map propagating light pulses into stationary spin excitations in atomic media^{13–16}. Here we describe and experimentally demonstrate a technique in which light propagating in a medium of Rb atoms is converted into an excitation with localized, stationary electromagnetic energy, which can be held and released after a controllable interval. Our method creates pulses of light with stationary envelopes bound to an atomic spin coherence, offering new possibilities for photon state manipulation and nonlinear optical processes at low light levels.

Several techniques to generate stationary light pulses in an atomic medium have recently been proposed theoretically^{17,18}. The present work is related conceptually to the ideas of ref. 18, but here we use a different approach that involves spatial modulation of the absorptive properties of the medium. The use of such a dissipative effect for coherent control of light is a novel phenomenon, closely related to earlier work on matched pulse propagation in electromagnetically induced transparency (EIT)^{2,19}. Although there exists a substantial body of work on localization of light in random dielectrics²⁰, Bragg gratings and photonic bandgap materials^{21–23}, the unique feature of the present method is that it allows accurate control of the creation of a stationary light pulse and its release.

The present method can be understood qualitatively by considering a three-state 'lambda' configuration of atomic states (Fig. 1a). A large ensemble of N atoms is initially prepared in the ground state, $|g\rangle$. We use forward and backward control beams, with time varying Rabi frequencies $\Omega_+(t)$ and $\Omega_-(t)$, respectively, to manipulate a weak pulse of signal light. In our experimental realization, the two control fields have identical frequencies but opposite propagation directions. The usual EIT¹⁰ corresponds to simultaneous propagation of the forward control and signal beams. When their frequency difference matches the level splitting between the ground state and a metastable ('spin-flipped') state ($|s\rangle$) the medium becomes transparent for the signal light, while the sharp atomic dispersion allows one to slow and localize an input signal pulse in the medium^{11,12}. By turning the control beam off while the pulse is in the medium¹³, the signal amplitude vanishes while its state is stored in a stationary spin coherence. This atomic excitation can be converted back into a light pulse, propagating in the forward or the backward direction, by application of the corresponding control beam^{14–16}. The atomic coherence can be converted into a stationary photonic excitation if the medium is illuminated simultaneously by forward and backward control beams. Specifically, if the two create a standing wave pattern, the EIT suppresses the signal absorption everywhere but in the nodes of the standing wave, resulting in a sharply peaked, periodic modulation of the atomic absorption for the signal light (Fig. 1b). Illumination by these beams also results in partial conversion of the stored atomic spin excitation into sinusoidally modulated signal light, but the latter cannot propagate in the medium owing to Bragg reflections off the sharp absorption peaks, resulting in vanishing group velocity of the signal pulse. Only after one of the control beams is turned off does the pulse acquire a finite velocity and can thus leave the medium in the direction of the remaining control beam.

To quantify these effects theoretically, we consider the interaction of atoms with resonant optical fields, represented by plane waves. We decompose the signal field into components propagating in the forward and backward directions along the z axis with wavevectors $\pm k$ and slowly varying amplitudes $E_{\pm}$. Following refs 13 and 18, we introduce two components $\Psi_{\pm}$ of a coupled excitation of light and an atomic spin wave ('dark-state polariton') corresponding to forward and backward signal fields, respectively. In the experimentally relevant case of small group velocities, the polariton components are represented by $\Psi_{\pm} = g\sqrt{NE_{\pm}}/\Omega_{\pm}$, where g is the atom–

field coupling constant¹. Assuming further slowly varying pulses and negligible spin decoherence, we find that the components evolve according to:

$$\frac{\partial}{\partial z} \Psi_+ = -\alpha_- \xi (\Psi_+ - \Psi_-) - \frac{1}{c} \frac{\partial}{\partial \tau} (\alpha_+ \Psi_+ + \alpha_- \Psi_-) \quad (1)$$

$$\frac{\partial}{\partial z} \Psi_- = -\alpha_+ \xi (\Psi_+ - \Psi_-) + \frac{1}{c} \frac{\partial}{\partial \tau} (\alpha_+ \Psi_+ + \alpha_- \Psi_-) \quad (2)$$

and the spin coherence $S = N^{-1/2}(\alpha_+ \Psi_+ + \alpha_- \Psi_-)$ with $\alpha_{\pm} = |\Omega_{\pm}|^2 / (|\Omega_+|^2 + |\Omega_-|^2)$. These equations describe two slow waves that are coupled owing to periodic modulation of atomic absorption and group velocity. The first term in the right-hand side of equations (1) and (2) is proportional to an absorption coefficient ξ near the resonant line centre (away from line centre ξ is imaginary and proportional to the dispersive phase shift of signal light). When ξ is large, this term gives rise to the pulse matching phenomenon^{2,19}: whenever one of the fields is created, the other will adjust itself within a short propagation distance to match its amplitude such that $\Psi_+ - \Psi_- \rightarrow 0$. The scaled time, $\tau(t) = \int_0^t dt (|\Omega_+|^2 + |\Omega_-|^2) / g^2 N$, reflects the group velocity reduction associated with atomic dispersion. Finally, the centre frequency of the signal light

was chosen to match its wavevector to that of the periodic absorption modulation. Owing to the steep atomic dispersion, this condition can be satisfied by a small detuning of the signal light, in direct analogy with other EIT-based processes².

In the case when the forward signal field propagates in the presence of only one (forward) control field, equation (1) has a simple solution $\Psi_+(z, t) = \Psi_+(z - c\tau(t), 0)$, which describes a coherent deceleration (Fig. 1c). Specifically, when the forward control beam is turned off, the polariton is stopped, the signal light vanishes, and a stationary spin excitation is created. Note that the spatial length of the spin excitation, l , is determined by the length of the compressed pulse corresponding to a product of the input pulse duration and the initial group velocity v_g^0 . Finally, when the forward control beam is turned back on, the propagating signal pulse is recreated.

When both control beams are turned on simultaneously, forward and backward signal components are both generated. The electric field amplitudes for both forward and backward signal light are proportional to the amplitudes of the corresponding control fields, as required by the pulse matching condition^{2,19}. Interesting insight into the propagation dynamics can be gained by considering the dispersion relation associated with equations (1) and (2), $\omega =$

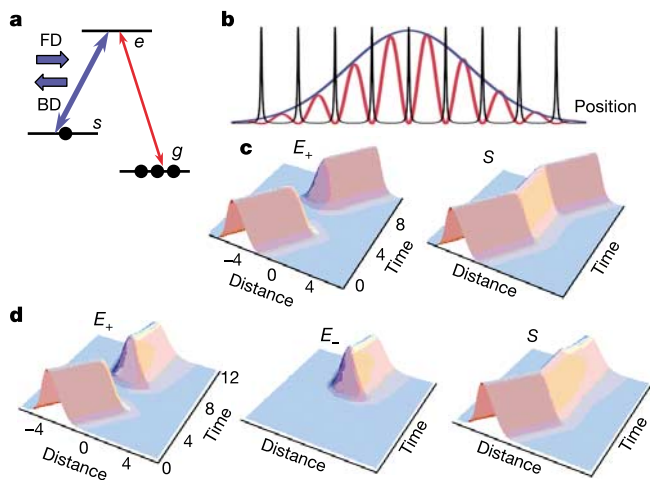


Figure 1 Physics of stationary pulses of light. **a**, Schematic of the atomic system. The signal beam (red arrow) is resonant with the transition from the ground to the excited state ($|e\rangle$) and the control fields (blue arrows: FD, forward; BD, backward) are tuned to resonance between the $|s\rangle$ state and the excited state. **b**, Schematic illustration of the spatial variation of the signal field absorption (black line), and the electric field (red line) in the stationary pulse in the medium along the direction of the control beam. Blue line represents initial atomic spin coherence. **c**, Storage of the weak signal pulse in Raman coherence. Calculated evolution is shown for the forward signal pulse and atomic spin wave amplitudes as a function of distance (in the units of compressed pulse length, l) and time (in units of l/v_g^0). While the forward pulse propagates the forward control beam is turned off (at the time $t = 4l/v_g^0$) and the pulse is stored in a stationary spin coherence. At the time $t = 8l/v_g^0$ the forward control field is turned back on, recreating the propagating pulse. Calculations are based on equations (1) and (2) with gaussian pulses and $\alpha_+ = 1, \alpha_- = 0$. **d**, Calculated evolution of the forward and backward signal components and atomic spin wave amplitude for the case when the stored coherence is illuminated (at time $t = 8l/v_g^0$) by forward and backward control beams with equal Rabi frequencies. Here stationary signal pulses are created. The main contributions to the dynamics include a slow spreading of the amplitudes as well as small shifts (of the order of $1/\xi$) of forward and backward components in the corresponding directions relative to the atomic coherence. Calculations are based on equations (1) and (2) with $\alpha_+ = \alpha_- = 1/2, \xi/l = 10$. All simulations assume that $v_g \ll c$, in which case the polaritons are mostly atomic and the amplitude of the spin coherence does not change significantly while the pulse is in the medium. Amplitude of fields and coherence are normalized to their initial value at $t = 0$.

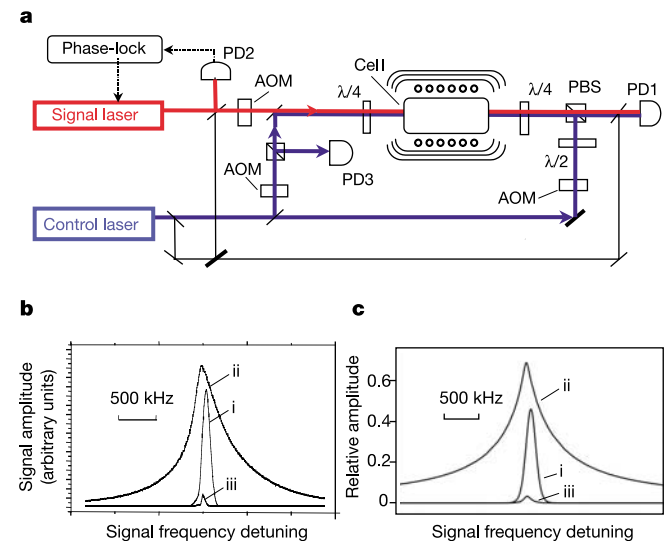


Figure 2 Experimental set-up, results of c.w. experiments and simulations. **a**, The signal field is generated by an extended cavity diode laser. The control beam from a locked Ti:sapphire laser is split to provide both forward and backward control fields. The signal laser is phase-locked to the control laser with a frequency offset corresponding to the hyperfine splitting of ^{87}Rb (6.84 GHz). All three fields have circular polarizations in the cell, but their intensities are controlled independently by acousto-optic modulators (AOMs). The beams overlap inside the cell at an angle of $\sim 10^{-4}$ rad. PD1 and PD3 are fast photodiodes used for heterodyne detection of the signal amplitude using a reference beam (black line). These are followed by a spectrum analyser running in zero span mode (the detection bandwidth is 3 MHz). The c.w. power of the signal beam is 250 μW , while the powers of the forward and backward control beams are independently adjusted between 8.0 mW and 40 mW. The beam size inside the cell is about 2 mm. $\lambda/2$ and $\lambda/4$ are half- and quarter-wave plates respectively. PBS, polarizing beam-splitter. **b**, Curve i represents the EIT signal transmission when the forward control beam is on and the backward control beam is off. Transmission is approximately 50% at EIT resonance. Curve ii is the reflected signal, and curve iii is the signal transmission when both forward and backward control fields are on. **c**, Corresponding theoretical simulations of the transmission and reflection of signal field from the medium composed of atoms as in Fig. 1a. Parameters correspond to atomic Rb, $\Omega_+ = \Omega_- = 15$ MHz, the spin decoherence rate of 3 kHz, and medium length corresponding to resonant attenuation of e^{-15} . The small frequency shift in the reflection resonance accounts for wavevector matching to satisfy the Bragg condition.

$-ck(\xi[\alpha_+ - \alpha_-] - ik)/(\xi - ik[\alpha_+ - \alpha_-])$, where ω is the Fourier frequency corresponding to the scaled time τ , and k is the spatial wavevector of the signal pulse envelopes. For small wavevectors and a large absorption coefficient, this corresponds to a linear dispersive medium with a group velocity $v_g = c(|\Omega_+|^2 - |\Omega_-|^2)/g^2N$. Thus, the time-domain solution describes the coupled motion of the $\Psi_{\pm}$ envelopes at a group velocity proportional to the intensity difference

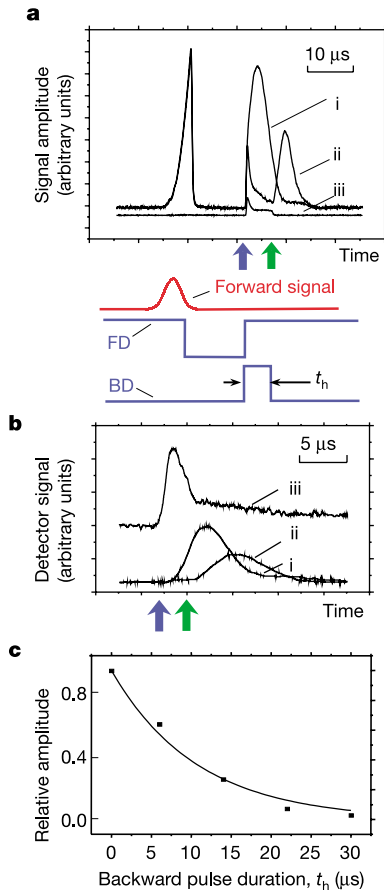


Figure 3 Results of pulsed experiments. **a**, Top, detected output signals from the Rb cell; bottom, timing diagram. Curve i is the signal detected in the forward direction resulting from pulse storage in a spin coherence. The left peak represents the fraction of the signal pulse that leaves the cell before the trapping has begun. (We were able to suppress this fraction to $\sim 10\%$.) It is delayed by $\sim 10\ \mu\text{s}$ compared to the original pulse (see timing diagram). The right peak in curve i appears only after the forward control is turned on, and thus represents the stored and retrieved signal. Curve ii is the same except the forward and backward control beams (FD and BD) are both turned on, as shown on the timing diagram and marked by the blue arrow. Curve iii is the signal in the backward direction under the same conditions. (This curve is plotted on a different scale, with the peak signal about a factor of five weaker than that on curve ii). The pulse is released in the forward direction when the backward control is turned off (green arrow). On the timing diagram, the rise time edges of the control pulses are about $0.1\ \mu\text{s}$. Note that the frequencies of the two control fields do not need to be exactly equal. This is in agreement with theory, and was verified by obtaining results similar to those in curve ii with backward control beam shifted by 80 MHz from the forward control. **b**, Rb fluorescence measured at the side of the cell. Curve iii is fluorescence associated with signal light during the release with both forward and backward control beams on. Background fluorescence associated with control beams is subtracted. Curves i and ii (shown for reference) correspond to the same signals as curves i and ii in **a**. Note that fluorescence measurements are carried out under conditions differing from those for other data (including **a**), because the photodetector inserted inside the magnetic shields introduces stray magnetic fields resulting in shortened spin coherence and storage times; this measurement is also detection-bandwidth limited. **c**, Dependence of the released signal pulse magnitude on the backward control pulse duration t_h .

between the two control fields. When the control Rabi frequencies are identical, the group velocity v_g vanishes and a stationary pulse of light is created (Fig. 1d). The electric field amplitudes of two signal components are then equal, $E_+ = E_-$, corresponding to sinusoidally modulated signal intensity, and the dispersion relation becomes $\omega = ick^2/\xi$. In the time domain, this corresponds to a slow spreading of the stationary pulse at a rate $\delta l/l \approx c\tau/(\xi l^2)$, which determines the maximal holding time. Hence, in an optically dense medium ($\xi l \gg 1$) a stationary photonic excitation can be controllably created.

Our experimental apparatus used to demonstrate this effect is shown in Fig. 2a. A magnetically shielded 4-cm-long cell containing ^{87}Rb vapour is maintained at temperature $T \approx 90^\circ\text{C}$ (atom number density is $10^{12} - 10^{13}\ \text{cm}^{-3}$). Long-lived hyperfine sublevels of the electronic ground state $S_{1/2}$ with $F = 1, 2$ are used as the storage states $|g\rangle$, $|s\rangle$, respectively, coupled via the excited state $P_{1/2}$. The hyperfine coherence time is limited by atomic diffusion out of the beam volume, and is enhanced with a Ne buffer gas at a pressure of 6 torr (resulting in spin-coherence lifetimes in the microsecond to millisecond range^{12,15}). Note that the Doppler shifts due to atomic motion do not directly affect the spin coherence when the control and the signal beams propagate collinearly in forward and backward pairs.

We first consider continuous wave (c.w.) excitation of the atomic Rb. Without the control beams, the atomic medium is completely opaque to the forward signal beam. When the forward control field is present, a sharp resonance, a few hundred kilohertz wide, appears in the transmission spectrum of the signal beam, corresponding to EIT (curve i in Fig. 2b). Turning on the backward control beam in the c.w. regime greatly reduces the EIT transmission (curve iii), while at the same time generating a reflected signal beam that is detected in the backward direction (curve ii). The peak reflection intensity is a substantial fraction (up to $\sim 80\%$) of the input signal beam.

These results demonstrate the possibility of coherent control of light via simultaneous driving of the medium with FD and BD control beams. Specifically, the signal light cannot propagate in the modulated EIT medium, but instead of absorbing the signal the medium reflects it as a high-quality Bragg mirror. This effect is analogous to that predicted theoretically¹⁸. However, the broad lineshape of curve ii indicates that periodic modulation (grating) of the absorptive rather than dispersive properties lies at the origin of the observed Bragg reflection²¹. We note, in particular, good agreement between the present experimental results and a theoretical model based on the resonant EIT medium with forward and backward control fields (Fig. 2c). At the same time, a qualitatively different lineshape is expected for dispersive Bragg gratings^{18,23}.

Turning to experiments with pulsed light, we first map the input signal pulse onto an atomic coherence of the Rb atoms^{14,15}. This procedure corresponds to Fig. 1c, and its experimental observation is shown by curve i in Fig. 3a. The atoms are first optically pumped into the lowest state. A gaussian-shaped signal pulse of about $5\ \mu\text{s}$ duration then enters the medium, where it is slowed to $v_g^0 \approx 6\ \text{km s}^{-1}$. The forward control beam is subsequently turned off. As a rule, a fraction of the signal pulse leaves the cell before that, leading to the first peak of curve i. When the forward control field is turned back on, the stored atomic excitation is converted back into light, which is detected in the forward direction (second peak of curve i). The amplitude of the retrieved light decays with increasing storage interval with a characteristic timescale of about $20\ \mu\text{s}$, corresponding to decay of the hyperfine coherence caused by atomic diffusion. Similar experimental results are obtained by detecting the signal light in the backward direction when the stored coherence is retrieved with the backward control beam¹⁶.

We next consider the retrieval of the atomic excitations by simultaneous application of the forward and backward control beams. When the intensities of the beams are carefully adjusted, the output signal pulses in both forward and backward directions

are greatly suppressed (curves ii and iii in Fig. 3a). Both channels exhibit small leakage. We attribute the first peak to photons retrieved near the cell boundaries, which do not experience sufficient Bragg reflections to be trapped efficiently. The long tail is probably due to a slow spreading of the stored pulse. When the backward beam is turned off, the released pulse is detected in the forward channel (curve ii). The presence of signal light inside the cell during the simultaneous application of the two control beams was verified directly by monitoring fluorescence from the side of the cell (Fig. 3b). For times when the signal output in the forward and backward directions is greatly suppressed, we observed significant enhancement of the signal light fluorescence (curve iii in Fig. 3b), due to residual atomic absorption.

These observations provide evidence for controlled conversion of the stored atomic coherence into a stationary photonic excitation in the cell. Note, in particular, that the magnitude of the fluorescence drops sharply after the backward pulse is turned off. This drop is followed by a gradual decay, associated with the exit of the slow pulse from the medium. This behaviour is in qualitative agreement with our simple model, which predicts the light intensity in the stationary pulses to be double that in the slowly propagating pulse. As shown in Fig. 3c, the magnitude of the released pulse decreases exponentially with increasing trapping time with a characteristic time constant of about 7 μ s. Note that only a part of this decay is due to the hyperfine coherence decay. Other decay mechanisms include spreading of the stationary pulse, as well as imperfect EIT. We anticipate that improvements in efficiency can probably be achieved by initial optical pumping into a single atomic sublevel, using an atomic system with larger level spacing or using the sharper absorption lines of cold atom clouds.

We finally outline possible avenues opened by the present work. First, we note that our procedure is based on a passive medium, and in the ideal limit is not accompanied by optical loss or gain, hence avoiding the associated noise. We therefore anticipate that our method preserves the quantum states of light pulses. (This is in contrast to Bragg gratings based on gain modulation²¹.) Second, although the present work demonstrates stationary light localization and storage in one dimension, it should be possible to controllably localize and guide stationary photonic pulses in three spatial dimensions by using control beams with properly designed wavefronts. Third, controlled conversion of propagating light into stationary light pulses opens interesting possibilities for enhanced nonlinear optical processes by combining the present technique with the resonant enhancement of nonlinear optics via EIT^{24–26}. This combination may enable controlled interactions involving quantum few-photon fields^{27–30} analogous to those feasible in cavity quantum electrodynamics⁵. Finally, extension of the present ideas to other systems might be possible using, for example, dynamic modulation of photonic bandgap materials. □

Received 6 September; accepted 4 November 2003; doi:10.1038/nature02176.

1. Scully, M. O. & Zubairy, S. M. *Quantum Optics* (Cambridge Univ. Press, Cambridge, UK, 1997).
2. Harris, S. E. Electromagnetically induced transparency. *Phys. Today* **50**(7), 36–42 (1997).
3. Lukin, M. D. & Imamoglu, A. Controlling photons using electromagnetically induced transparency. *Nature* **413**, 273–276 (2001).
4. Matsko, A. B. *et al.* Slow, ultraslow, stored, and frozen light. *Adv. At. Mol. Opt. Phys.* **46**, 191–242 (2001).
5. Mabuchi, H. & Doherty, A. C. Cavity quantum electrodynamics: Coherence in context. *Science* **298**, 1372–1377 (2002).
6. Cirac, J. I. *et al.* Quantum state transfer and entanglement distribution among distant nodes in a quantum network. *Phys. Rev. Lett.* **78**, 3221–3224 (1997).
7. Kuzmich, A. *et al.* Generation of nonclassical photon pairs for scalable quantum communication with atomic ensembles. *Nature* **423**, 731–734 (2003).
8. van der Wal, C. H. *et al.* Atomic memory for correlated photon states. *Science* **301**, 196–200 (2003).
9. Boyd, R. W. *Nonlinear Optics* (Academic, New York, 1992).
10. Boller, K. J., Imamoglu, A. & Harris, S. E. Observation of electromagnetically induced transparency. *Phys. Rev. Lett.* **66**, 2593–2596 (1991).
11. Hau, L. V. *et al.* Light speed reduction to 17 metres per second in an ultracold atomic gas. *Nature* **397**, 594–598 (1999).
12. Kash, M. M. *et al.* Ultraslow group velocity and enhanced nonlinear optical effects in a coherently driven hot atomic gas. *Phys. Rev. Lett.* **82**, 5229–5232 (1999).
13. Fleischhauer, M. & Lukin, M. D. Dark-state polaritons in electromagnetically induced transparency.

Phys. Rev. Lett. **84**, 5094–5097 (2000).

14. Liu, C. *et al.* Observation of coherent optical information storage in an atomic medium using halted light pulses. *Nature* **409**, 490–493 (2001).
15. Phillips, D. F. *et al.* Storage of light in atomic vapor. *Phys. Rev. Lett.* **86**, 783–786 (2001).
16. Zibrov, A. S. *et al.* Transporting and time reversing light via atomic coherence. *Phys. Rev. Lett.* **88**, 103601 (2002).
17. Kocharovskaya, O., Rostovtsev, Y. & Scully, M. O. Stopping light via hot atoms. *Phys. Rev. Lett.* **86**, 628–631 (2001).
18. Andre, A. & Lukin, M. D. Manipulating light pulses via dynamically controlled photonic band gas. *Phys. Rev. Lett.* **89**, 143602 (2002).
19. Harris, S. E. Electromagnetically induced transparency with matched pulses. *Phys. Rev. Lett.* **70**, 552–555 (1993).
20. Cao, H. *et al.* Random laser action in semiconductor powder. *Phys. Rev. Lett.* **82**, 2278–2281 (1999).
21. Kogelnik, H. & Shank, C. V. Coupled wave theory of distributed feedback lasers. *J. Appl. Phys.* **43**, 2327–2335 (1972).
22. Yablonovich, E. *et al.* Donor and acceptor modes in photonic bandgap structures. *Phys. Rev. Lett.* **67**, 3380–3383 (1991).
23. Slusher, D. & Eggleston, B. (eds) *Nonlinear Photonic Crystals* (Springer, New York, 2003).
24. Harris, S. E., Field, J. E. & Imamoglu, A. Nonlinear optical processes using electromagnetically induced transparency. *Phys. Rev. Lett.* **64**, 1107–1110 (1990).
25. Schmidt, H. & Imamoglu, A. Giant Kerr nonlinearity obtained by electromagnetically induced transparency. *Opt. Lett.* **21**, 1936–1938 (1996).
26. Hemmer, P. R. *et al.* Efficient low-intensity optical-phase conjugation based on coherent population trapping in sodium. *Opt. Lett.* **20**, 982–984 (1995).
27. Harris, S. E. & Hau, L. V. Nonlinear optics at low light levels. *Phys. Rev. Lett.* **82**, 4611–4614 (1999).
28. Lukin, M. D. & Imamoglu, A. Nonlinear optics and quantum entanglement of ultraslow single photons. *Phys. Rev. Lett.* **84**, 1419–1422 (2000).
29. Petrosyan, D. & Kurizki, G. Symmetric photon-photon coupling by atoms with Zeeman-split sublevels. *Phys. Rev. A* **65**, 033833 (2002).
30. Ottaviani, C., Vitali, D., Artoni, M., Cataliotti, F. & Tombesi, P. Polarization qubit phase gate in driven atomic media. *Phys. Rev. Lett.* **90**, 197902 (2003).

Acknowledgements We thank A. Andre, M. Eisaman, L. Childress, C. van der Wal, R. Walsworth, S. Zibrov and T. Zibrova for discussions, experimental help and comments on the manuscript. This work is supported by the NSF, the DARPA, the David and Lucille Packard Foundation and the Alfred Sloan Foundation. Partial support by the ONR (DURIP) is also acknowledged.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.D.L. (lukin@physics.harvard.edu).

An off-normal fibre-like texture in thin films on single-crystal substrates

C. Detavernier^{1,2}, A. S. Özcan³, J. Jordan-Sweet¹, E. A. Stach⁴, J. Tersoff¹, F. M. Ross¹ & C. Lavoie¹

¹IBM T. J. Watson Research Center, Yorktown Heights, New York, New York 10598, USA

²Department of Solid State Physics, Ghent University, 9000 Ghent, Belgium

³Department of Physics, Boston University, Boston, Massachusetts 02215, USA

⁴National Center for Electron Microscopy, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

In the context of materials science, texture describes the statistical distribution of grain orientations. It is an important characteristic of the microstructure of polycrystalline films^{1–5}, determining various electrical, magnetic and mechanical properties. Three types of texture component are usually distinguished in thin films: random texture, when grains have no preferred orientation; fibre texture^{6–10}, for which one crystallographic axis of the film is parallel to the substrate normal, while there is a rotational degree of freedom around the fibre axis; and epitaxial alignment (or in-plane texture) on single-crystal substrates^{11–15}, where an in-plane alignment fixes all three axes of the grain with respect to the substrate. Here we report a fourth type of texture—which we call axiotaxy—identified from complex but

symmetrical patterns of lines on diffraction pole figures for thin films formed by solid-state reactions. The texture is characterized by the alignment of planes in the film and substrate that share the same d -spacing. This preferred alignment of planes across the interface manifests itself as a fibre texture lying off-normal to the sample surface, with the fibre axis perpendicular to certain planes in the substrate. This texture forms because it results in an interface, which is periodic in one dimension, preserved independently of interfacial curvature. This new type of preferred orientation may be the dominant type of texture for a wide class of materials and crystal structures.

Texture is frequently studied using X-ray diffraction (XRD). A pole figure for a particular (hkl) plane depicts the statistical angular distribution of the direction of the normal to this plane^{1–3}. One can imagine placing the sample in the centre of an imaginary hemisphere and, for each grain in the film, marking the intersection between the normal to the chosen (hkl) plane and the sphere. By projecting the density of marks on the sphere onto a planar surface, one obtains the (hkl) pole figure (Fig. 1a). The angles ϕ and χ are defined as spherical coordinates (azimuth and elevation) in the hemisphere.

Pole figures for thin films typically consist of simple geometric features (Fig. 1b, c). Random texture results in featureless pole figures. Fibre texture^{4–10} (Fig. 1b) can be identified by the occurrence of circles centred around the middle of the pole figure. (For some materials, the fibre axis is at a small angle to the surface normal^{16–18}; this variant of fibre texture is sometimes referred to as tilted texture.) Epitaxial alignment (also referred to as ‘in-plane alignment’ or ‘in-plane texture’)^{11–15} can be identified by the presence of well-defined spots at certain locations on the pole figure (Fig. 1c). The three types of texture (random, fibre texture and epitaxial) are frequently observed simultaneously within the same film. For instance, multiple variants of in-plane alignment can coexist within a CoSi_2 film on Si(100) (refs 11, 12). When different texture

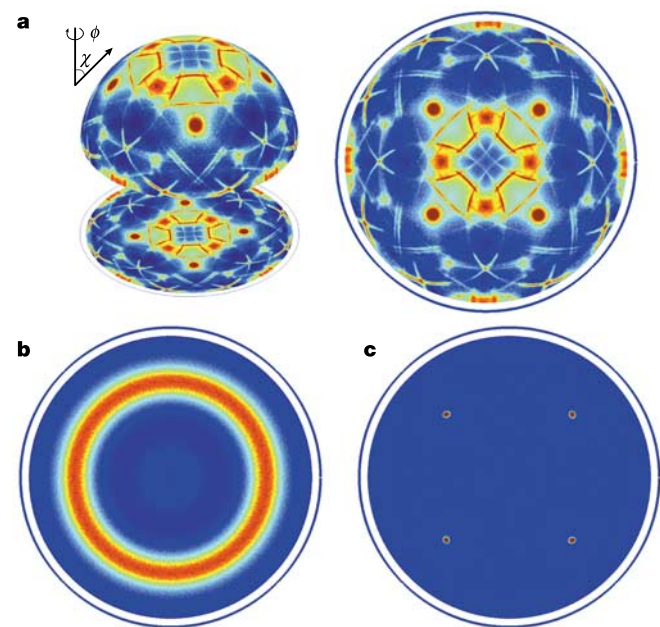


Figure 1 Comparison with standard thin-film pole figures. **a**, Illustration of the projection method (left) used to obtain the NiSi(112) pole figure (right) for a 60-nm NiSi layer formed by annealing in ultrahigh vacuum. **b**, Example of ‘standard’ fibre texture: a Cu(200) pole figure illustrating (111)-fibre texture for a 500-nm-thick film of Cu sputter deposited on SiO_2 . **c**, Example of ‘standard’ epitaxial alignment: a Cu(111) pole figure showing the epitaxial alignment of a 500-nm Cu film deposited on HF-cleaned Si(001). Regions of high (low) diffracted intensity are coloured red (blue).

components are present within the same film, each of these texture components will contribute to the pole figures, which will then consist of different sets of rings and/or spots.

Here we report on intriguingly complex pole figures containing symmetrical line patterns that were measured for several materials formed by a solid-state reaction on a single-crystal substrate. These lines, evident in the spherical representation of the NiSi(112) pole figure in Fig. 1a, are shown in more detail in Fig. 2a–c, which illustrates three pole figures for a 60-nm NiSi film on Si(001). The figures represent the raw data, without any smoothing, fitting, noise reduction or other data enhancement techniques. The patterns of lines cannot be described by either fibre texture (which would result in concentric circles), epitaxial alignment (which would result in spots) or by any combination of a discrete set of known types of texture. Therefore, these patterns are characteristic of a new type of texture, for which we propose the name *axiotaxy*. Although we only show results here for a 60-nm-thick film of orthorhombic NiSi on Si(001), the occurrence of axiotaxy appears to be a robust phenomenon. It is determined primarily by the selection of film and

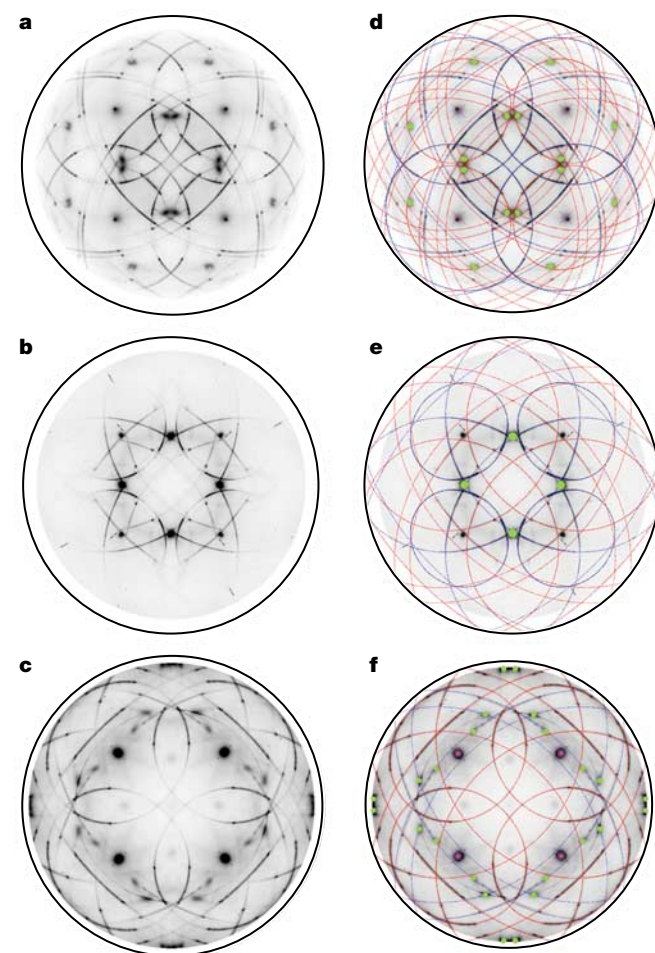


Figure 2 Pole figures and calculated pattern of lines for a NiSi film on Si(001). **a–c**, Pole figures for a 60-nm NiSi film on Si(001); **a**, (112); **b**, (103); and **c**, (202)/(211). The film was formed by annealing in nitrogen for 30 s at 550 °C. The diffracted intensity is shown on a linear grey-scale ranging from 0 (white) to 3,000 (black) counts above the background. **d–f**, The same data with calculated line patterns overlaid on top of the measurement. These patterns were calculated assuming that NiSi(211) (red lines) or (202) planes (blue lines) are parallel to Si(110)-type planes in the substrate (that is, fibre axes at $\chi = 45^\circ$, $\phi = 45^\circ, 135^\circ, 225^\circ, 315^\circ$). Green dots, the calculated pattern for the epitaxial texture component (see text); magenta crosses, location of signal from Si substrate peaks.

substrate material, and to first order is independent of processing conditions. Indeed, we observed similar features for NiSi films on Si(001), (111) or (110) substrates with film thickness ranging from 16 to 200 nm, annealed in nitrogen, purified helium or ultrahigh vacuum at temperatures ranging from 400 to 750 °C. This indicates that the observations reported here are not related to specific annealing conditions or the presence of impurities. The solid-state reaction between the deposited Ni film and the Si substrate is known to result in a phase sequence, with Ni first transforming into metal-rich silicides, followed by transformation into NiSi at higher temperature by continued Ni diffusion. It could be thought that the observed texture of NiSi is related to this specific formation sequence, owing to (for example) some particular property of the precursor phase. However, we observed similarly complex patterns on pole figures for other films also formed by solid-state reaction (tetragonal α -FeSi₂/Si(001), cubic CoSi₂/Si(001), orthorhombic NiGe/Ge(001) and orthorhombic C54-TiSi₂/Si(001)), indicating that the phenomenon of axiotaxy is not related to some peculiarity of a single material (for example, the texture of a specific precursor phase, or a diffusion- versus nucleation-controlled formation process) or even of a single class of crystal structures (Supplementary Figs A1 and A2).

The pole figures (Fig. 2) allow unambiguous identification of the preferential orientation of the grains with respect to the reference frame provided by the single-crystal substrate. The features appear circular on the spherical representation in Fig. 1a, so this suggests that the patterns of lines are related to a fibre-like texture with the fibre axes at $\chi = 45^\circ$ (instead of normal to the substrate—that is, at $\chi = 0^\circ$ —as for standard fibre texture). Indeed, we found that the complex patterns on Figs 1a and 2 are generated by two off-normal fibre textures, each with fibre axes at $\chi = 45^\circ$ and four symmetry-related ϕ values ($45^\circ, 135^\circ, 225^\circ, 315^\circ$). As the fibre axes are aligned along Si(110)-type directions in the substrate, the preferential orientation of the grains that are part of the axiotaxy components

is defined by the constraint of aligning NiSi (202) or (211) planes in the film parallel to Si(110)-type planes in the substrate.

The intricate geometrical patterns in the data (Fig. 2a–c) are reproduced in Fig. 2d–f with calculated patterns for the three main texture components overlaid on the measured patterns, illustrating the excellent correspondence. The calculations should not be considered as curve fitting because they are based purely on geometry, without any adjustable parameters. In addition to the two axiotaxy components, the pole figures contain intense spots that are not part of the lines. These spots are caused by a standard epitaxial alignment, for which the NiSi(200) plane is almost parallel to Si(010) and the NiSi(014) plane is parallel to Si(001). A transmission electron microscopy (TEM) cross-section for such a grain is shown in Supplementary Fig. A3.

We characterized the morphology of the NiSi grains using cross-sectional TEM analysis. The film consists of a single layer of grains typically 100 nm across (Fig. 3a). The interface between the NiSi and the substrate is abrupt (that is, no amorphous interlayer) and not faceted. Considerable grain boundary grooving was observed, indicating that for this particular annealing temperature, agglomeration had already started, and the NiSi/Si interface was therefore not flat. The high-resolution TEM micrograph of Fig. 3b provides direct corroborating evidence for the preferred alignment of NiSi(202) or (211) planes parallel to Si(220) planes. To show this more clearly, the image was filtered to keep only intensity associated with the NiSi(211)/(202) and Si(110) periodicity (Fig. 3c), confirming the alignment of the planes in the film and substrate.

Although an understanding of the crystallographic orientation of

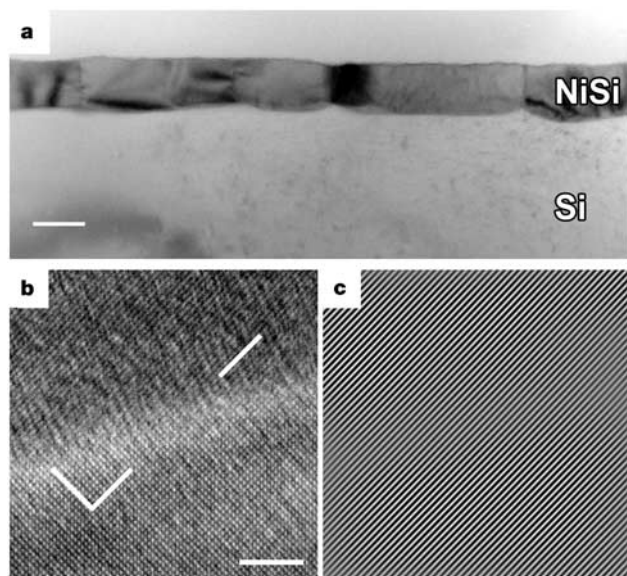


Figure 3 Transmission electron micrographs of axiotaxial NiSi grains. **a**, Low-magnification image showing the grain structure. The typical grain size is 100 nm, and some grain boundary grooving is visible. Scale bar, 50 nm. **b**, Higher-resolution image showing the correspondence of lattice planes in the Si and NiSi. The image has been Wiener filtered to slightly enhance lattice fringe contrast²⁰. The orientations of the NiSi and Si are shown, and the alignment of the lattice planes across the interface is indicated. Scale bar, 2.5 nm. **c**, A filtered version of **b** where all intensity is removed from the image apart from that associated with the NiSi(211)/(202) and Si(110) periodicity. This filtering allows us to confirm visually the alignment of these planes.

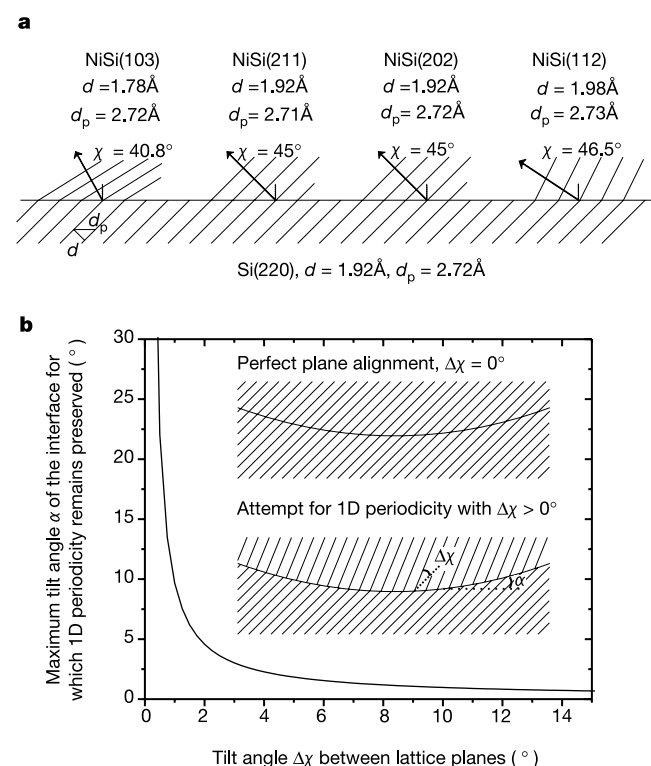


Figure 4 Illustration of the importance of plane alignment across the interface. **a**, A 1D periodic interface can be achieved by compensating the mismatch in d -spacing by tilting the lattice planes over an angle $\Delta\chi$, thus reducing the mismatch in projected d -spacing, Δd_p . However, this works only for a flat interface (insets in **b**). **b**, The graph shows the maximum angle α through which the interface can curve while maintaining good 1D periodicity (defined as mismatch $\Delta d_p < 0.5\%$), as a function of the tilt angle $\Delta\chi$. The larger the allowed value of α , the more stable the 1D periodic nature of the interface with respect to interfacial curvature.

the textured grains is interesting in itself, one is left to wonder why plane alignment across the interface is preferred, and why specifically the NiSi(202) and (211) planes are selected. We note that the interplanar spacings of 1.921 and 1.919 Å for NiSi(202) and (211) planes, respectively¹⁹, are within 0.06% of the spacing of 1.9201 Å for Si(220) planes, and the alignment of planes with almost identical *d*-spacing across the interface clearly results in a structure that is periodic along one direction in the plane of the interface (Fig. 3b).

However, unlike fibre texture and epitaxy, for which a single (*hkl*) plane is preferentially parallel to the interface, the interfaces of grains belonging to the same axiotaxy component consist of different crystallographic planes, depending on the rotation around the off-normal fibre axis. Any such rotation produces an interface that remains periodic along at least one direction, but some rotations will result in a higher degree of interfacial periodicity. In fact, we observe that the diffracted intensity varies along the arcs on the pole figures. We calculate that the most intense spots on the lines are caused by grain orientations for which periodicity is achieved along two independent directions in the plane of the interface (Supplementary Fig. A4).

Axiotaxy, resulting in a one-dimensional (1D) periodic interface with a possibly lower-energy bond configuration, can thus be considered as an intermediate case between a random interface (no periodicity) and epitaxy (a two-dimensional, 2D, periodic interface). However, it is important to note that although 1D periodicity can in principle be achieved for any combination of substrate and film planes by choosing an appropriate tilt angle $\Delta\chi$ (Fig. 4a shows some options for different NiSi planes), axiotaxy is experimentally observed only for 1D periodic interface structures that result from plane alignment across the interface (that is, with $\Delta\chi \approx 0^\circ$). To explain the apparent importance of plane alignment, we consider the effect of interfacial roughness (Fig. 4b insets). The interface of thin films is rarely perfectly flat, especially for films formed by a solid-state reaction. Suppose that at one point on a curved grain there is good 1D periodicity (0% mismatch) because of a properly chosen tilt angle $\Delta\chi$. Because of interface curvature, the projected *d*-spacing changes as one moves away from this point, reducing the quality of the 1D periodic matching (lower inset). One can calculate the mismatch in projected *d*-spacing Δd_p along the interface as a function of the tilt angle $\Delta\chi$ and an angle α that reflects the curvature of the interface (Fig. 4b, and Supplementary Fig. A5). For perfect plane alignment across the interface, $\Delta\chi = 0^\circ$ and α can be arbitrarily large, indicating that the 1D periodic nature of the interface remains preserved irrespective of interfacial roughness (upper inset). However, if $\Delta\chi$ is not zero, then as the interface curves away from a point with zero mismatch, the mismatch increases. If we select a mismatch of <0.5% as our criterion for good periodicity, Fig. 4b shows that for increasing $\Delta\chi$, the range of curvature over which the interface maintains good periodicity shrinks dramatically. We may conclude that (nearly) perfect alignment of planes in film and substrate that share (nearly) the same *d*-spacing is the only way in which a 1D periodic interface structure can be achieved, the periodic nature of which is preserved independently of interfacial curvature.

To corroborate the above model, we consider two weak axiotaxy-related texture components that are observed in the pole figures for the NiSi film when plotting the data using a logarithmic intensity scale (Supplementary Figs A6 and A7). These two additional components are caused by off-normal fibre texture related to NiSi(103) and (112) planes for which the fibre axis is at $\chi = 40.8^\circ$ and $\chi = 46.5^\circ$, respectively, and $\phi = 45^\circ, 135^\circ, 225^\circ, 315^\circ$. NiSi(103) has a *d*-spacing 7.6% smaller than Si(220), while NiSi(112) has a *d*-spacing 3% larger than Si(220), and this mismatch has been compensated by slightly tilting the NiSi planes (Fig. 4a). Indeed, the measured tilts of $\Delta\chi = -4.2^\circ$ for NiSi(103) and $\Delta\chi = +1.5^\circ$ for NiSi(112) reduce the mismatch Δd_p in *d*-spacing projected onto the interface to only -0.01% and 0.4% , respectively. This illustrates the

importance of achieving a 1D periodic interface structure (with $\Delta d_p \approx 0\%$) in providing the driving force for this type of off-normal fibre texture. However, the relatively weak and broad intensities of these two axiotaxy components, compared with the NiSi(202) or (211) components, corroborate the importance of plane alignment, as explained by the model.

We finally comment that the occurrence of axiotaxy may be controlled by adjusting the interplanar spacing of lattice planes (Supplementary Fig. A8). Adding Pt to the Ni–Si reaction is known to result in the formation of a solid solution with a NiSi-type structure with a larger unit cell than pure NiSi. Alloying with 5–10% Pt causes the (112)-related axiotaxy to disappear (as the increase in *d*-spacing would require an even higher tilt angle) and the (202)/(211)-related axiotaxy to weaken, while the (103) axiotaxy component becomes more defined and stronger in intensity as the (103) plane spacing increases and the planes become more parallel to Si(220) (that is, $\Delta\chi$ decreases). □

Methods

Sample preparation

Films of 8–100 nm of Ni, 30 nm of Ni(5% Pt), Fe, Co and Co(5% Ti) were sputter deposited onto HF-cleaned Si(001) substrates in an ultrahigh-vacuum deposition system. The samples were annealed in nitrogen ambient for 30 s at 950 °C (for Fe and Co) or 550 °C (for Ni). In addition, a 60-nm-thick NiSi film was formed by annealing *in situ*—that is, in ultrahigh vacuum (pressure 7×10^{-10} torr at the start of the anneal, heater held at 550 °C for 5 min, pressure increased to 1.6×10^{-9} torr at the end of the anneal). Standard $\theta/2\theta$ X-ray diffraction was used to confirm that annealing resulted in the formation of α -FeSi₂, CoSi₂ or NiSi, respectively, with no evidence of other phases present in the films.

Experimental procedure

Pole figures were measured at the X20A beamline of the National Synchrotron Light Source at Brookhaven National Laboratory. A Si monochromator was used to select a wavelength of 1.5406 Å. This wavelength was chosen to simplify comparison with laboratory-based measurements using CuK α radiation. The sample was mounted on a four-circle diffractometer (Schultz geometry). We used a scintillation counter to detect the diffracted intensity. By fixing the sample and detector at a given θ and 2θ angle, one fixes the *d*-spacing of the crystallographic plane in the film for which diffraction will be detected. The pole figure is then obtained by rotating the sample around the axis that is normal to its surface (ϕ scan) and around the axis that is formed by the intersection of the sample surface and the plane defined by the X-ray beam and detector (χ scan). The pole figures were acquired in steps of 0.5° in ϕ and χ ($0^\circ \leq \phi \leq 90^\circ$ and $0 \leq \chi \leq 85^\circ$). ϕ and χ alignment was achieved by setting ϕ and χ equal to 45° at the location of the Si(011) substrate peak. As a consequence, the directions ($\phi = 0^\circ, \chi = 0^\circ$), ($\phi = 0^\circ, \chi = 90^\circ$) and ($\phi = 90^\circ, \chi = 90^\circ$) correspond to the directions of the normal to the Si(001), (110) and (110) planes, respectively. Transmission electron microscopy was carried out at 800 kV accelerating voltage on the Atomic Resolution Microscope (which has a point resolution of 0.16 nm) at the National Center for Electron Microscopy, Lawrence Berkeley National Laboratory.

Received 20 March; accepted 4 November 2003; doi:10.1038/nature02198.

1. Kocks, U. F., Tome, C. N. & Wenk, H.-R. *Texture and Anisotropy: Preferred Orientations in Polycrystals and their Effect on Materials Properties* (Cambridge Univ. Press, Cambridge, UK, 1976).
2. Randle, V. & Engler, O. *Introduction to Texture Analysis: Microtexture, Microtexture and Orientation Mapping* (Gordon and Breach Science, Amsterdam, 2000).
3. Bunge, H. J. *Texture Analysis in Materials Science—Mathematical Methods* 2nd edn (Butterworths, London, 1982).
4. Nokinov, V. *Grain Growth and Control of Microstructure and Texture in Polycrystalline Materials* (CRC, Boca Raton, 1997).
5. Thompson, C. V. & Carel, R. Texture development in polycrystalline thin films. *Mater. Sci. Eng. B* **32**, 211–219 (1995).
6. Harper, J. M. E. & Rodbell, K. P. Microstructure control in semiconductor metallization. *J. Vac. Sci. Technol. B* **15**, 763–779 (1997).
7. Dong, L. & Srolovitz, D. J. Texture development mechanisms in ion beam assisted deposition. *J. Appl. Phys.* **84**, 5261–5269 (1998).
8. Murray, C. E. & Rodbell, K. P. Texture inheritance in Al(Cu) interconnect materials. *J. Appl. Phys.* **89**, 2337–2342 (2001).
9. Lucadamo, G., Barmak, K. & Rodbell, K. P. Texture in Ti/Al and Nb/Al multilayer thin films: role of Cu. *J. Mater. Res.* **16**, 1449–1459 (2001).
10. Schell, N., Matz, W., Bottiger, J., Chevallier, J. & Kringhoj, P. Development of texture in TiN films by use of *in situ* synchrotron x-ray scattering. *J. Appl. Phys.* **91**, 2037–2044 (2002).
11. Bulle-Lieuwma, C. W. T., van Ommen, A. H., Hornstra, J. & Aussems, C. N. A. M. Observation and analysis of epitaxial growth of CoSi₂ on (100) Si. *J. Appl. Phys.* **71**, 2211–2224 (1992).
12. Bulle-Lieuwma, C. W. T. Epitaxial growth of CoSi₂/Si structures. *Appl. Surf. Sci.* **68**, 1–18 (1993).
13. Harper, J. M. E., Rodbell, K. P., Colgan, E. G. & Hammond, R. H. Control of in plane texture of body centered cubic metal thin films. *J. Appl. Phys.* **82**, 4319–4326 (1997).
14. Dong, L., Srolovitz, D. J., Was, G. S., Zhao, Q. & Rollett, A. D. Combined out of plane and in plane texture control in thin films using ion beam assisted deposition. *J. Mater. Res.* **16**, 210–216 (2001).

15. Özcan, A. S. *et al.* Texture of TiSi₂ thin films on Si(001). *J. Appl. Phys.* **92**, 5011–5018 (2002).
16. Tsukada, M. & Ohfuji, S. Structural inheritance from polycrystalline underlayers in the growth of double-layered aluminum films. *J. Vac. Sci. Technol. B* **11**, 326–332 (1993).
17. Tracy, D. P., Knorr, D. B. & Rodbell, K. P. Texture in multilayer metallization structures. *J. Appl. Phys.* **76**, 2671–2680 (1994).
18. Knorr, D. B., Merchant, S. M. & Biberger, M. A. Development of texture in interconnect thin film stacks. *J. Vac. Sci. Technol. B* **16**, 2734–2744 (1998).
19. Joint Committee on Powder Diffraction Standards. *Powder Diffraction File* 38–0844 (International Centre for Diffraction Data, Philadelphia, 1998).
20. Kilaas, R. Optimal and near-optimal filter in high-resolution electron microscopy. *J. Microsc.* **190**, 45–51 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank F. M. d’Heurle, J. M. E. Harper, C. Cabral, K. Ludwig, C. Murray, C. Noyan and K. P. Rodbell for discussions, and C. Diareo for TEM sample preparation. C.D. thanks the ‘Fonds voor Wetenschappelijk Onderzoek—Vlaanderen’ for support. The work at NCEM was supported by the Director, Office of Science, Office of Basic Energy Sciences, Division of Materials Sciences and Engineering, of the US Department of Energy. Pole figures were measured at the NSLS, Brookhaven National Laboratory, which is supported by the US Department of Energy, Division of Materials Sciences and Division of Chemical Sciences.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.D. (christophe.detavernier@ugent.be) or C.L. (clavoie@us.ibm.com).

Coupled spatial variations in precipitation and long-term erosion rates across the Washington Cascades

Peter W. Reiners¹, Todd A. Ehlers², Sara G. Mitchell³ & David R. Montgomery³

¹Department of Geology and Geophysics, Yale University, 210 Whitney Avenue, New Haven, Connecticut 06511, USA

²Department of Geological Sciences, University of Michigan, Ann Arbor, Michigan 48109, USA

³Department of Earth and Space Sciences, University of Washington, Seattle, Washington 98195, USA

Past studies of tectonically active mountain ranges have suggested strong coupling and feedbacks between climate, tectonics and topography^{1–5}. For example, rock uplift generates topographic relief, thereby enhancing precipitation, which focuses erosion and in turn influences rates and spatial patterns of further rock uplift. Although theoretical links between climate, erosion and uplift have received much attention^{2,6–10}, few studies have shown convincing correlations between observable indices of these processes on mountain-range scales^{11,12}. Here we show that strongly varying long-term (>10⁶–10⁷ yr) erosion rates inferred from apatite (U–Th)/He cooling ages across the Cascades mountains of Washington state closely track modern mean annual precipitation rates. Erosion and precipitation rates vary over an order of magnitude across the range with maxima of 0.33 mm yr^{–1} and 3.5 mm yr^{–1}, respectively, with both maxima located 50 km west (windward) of the topographic crest of the range. These data demonstrate a strong coupling between precipitation and long-term erosion rates on the mountain-range scale. If the range is currently in topographic steady state, rock uplift on the west flank is three to ten times faster than elsewhere in the range, possibly in response to climatically focused erosion.

The Washington Cascades provide a natural laboratory for

observing interactions between climate, tectonics and topography. Although isolated volcanoes in the mountain range are part of an active magmatic arc extending south into California, most of the topographic relief in the Washington Cascades, and essentially all of it north of about 47°N, is the result of deep-seated bedrock uplift and the resultant erosion, rather than volcanic extrusion. Summits in the range seldom reach elevations higher than 2.7 km, but local relief is commonly 1.2–1.8 km. The modern Washington Cascades cast a dramatic orographic rain shadow. Mean annual precipitation rates on the windward west flank are as high as 4 mm yr^{–1} and host lush vegetation, in contrast to rates of 0.2 mm yr^{–1} or less in the sagebrush steppe directly east of the range. Palaeobotanical evidence¹³ indicates that the orographic rain shadow was not as strong before the Late Miocene epoch, so it is likely that at least some of the current topographic expression of the Washington Cascades is younger than 10–15 Myr. This is consistent with the well-known post-15-Myr BP warping and uplift of the lava flows of the east-derived Columbia River basalt group on the eastern side of the range. Little is known about either the kinematics or dynamics of uplift in the Washington Cascades. Global positioning system (GPS) data suggest only weak horizontal crustal shortening across the range¹⁴, and patterns of crustal seismicity and detailed geologic mapping do not suggest any large active Neogene faults or other structures (see the catalogued seismicity record at (<http://www.seismo.berkeley.edu/seismo/>); also available in the Supplementary Information)^{15–19}.

We measured bedrock apatite (U–Th)/He ages from samples in a broad (200 km) east–west swath across the Washington Cascades to examine spatial patterns of erosion in a setting with a strong orographic precipitation gradient. The apatite (U–Th)/He system has a closure temperature of ~60–70 °C (ref. 20) and ages generally represent the time since a rock passed through a depth in the crust corresponding to that temperature (typically 1.5–2.5 km), as a result of tectonic or erosional exhumation.

Thirty new apatite He ages from twenty samples from the western flank of the Cascades are shown in Fig. 1, along with our existing data²¹ from elsewhere in this swath. Samples were collected from plutons with crystallization ages ranging from ~20–95 Myr BP. In all cases except that of Mt Pilchuck in the far west, apatite He ages are much less than crystallization ages. In this part of the Cascades, plutons younger than ~20 Myr BP are small and rare and there is no evidence in age and pluton distributions that ages reflect a significant component of magmatically controlled heating.

Apatite He ages in this swath of the central Cascades vary from 4.4 to 60 Myr BP (Fig. 1). In single vertical transects collected over short horizontal distances, ages generally show increasing ages, or little change in age, with increasing elevation²¹. The most obvious pattern in the data consists of relatively old ages (>25 Myr BP) at the topographic crest and far eastern and western flanks of the range, and young ages (<12 Myr BP) on the west slope (Figs 1 and 2).

To first order, the ratio of the apatite He closure isotherm depth (~1.5–2.5 km) and age yields an estimate of time-averaged exhumation rate. As there are no known late Tertiary faults or extensional structures in this region^{15–19}, we assume that this exhumation is entirely erosional. We calculated erosion rates for these samples by relating each apatite He age to cooling-rate-dependent closure temperature and depth through a series of equations with assumed parameters, including geothermal gradient, He diffusion properties²⁰, thermal diffusivity, and depth to constant temperature²² (see Supplementary Information). We used sample-specific geothermal gradients based on interpolation of observed gradients ranging from 20 to 40 °C km^{–1} (ref. 23) (Fig. 1). Our model assumes a steady-state distribution of isotherms in the crust, which is fairly robust for the relatively slow erosion rates inferred from these data and tectonic setting over the past 40 Myr (ref. 23).

Closure depth for each sample was adjusted to account for local

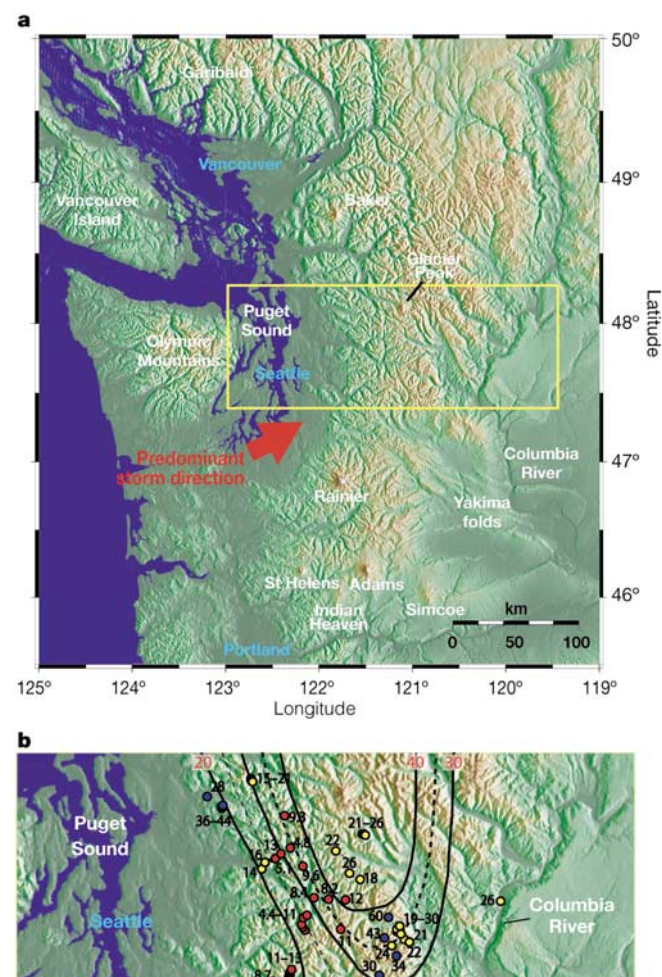


Figure 1 Map of the Washington Cascades and regional features, showing locations of samples and swath for climatic and topographic comparisons. White text, natural features including five active or recently active volcanoes; blue text, cities. Yellow box in **a** indicates location of expanded panel (**b**). **b**, Sample locations and ages (or age ranges, for multiple samples in vertical transects). Sample locations in red denote ages less than 13 Myr BP; yellow, 13–26 Myr BP; blue, >26 Myr BP. Contours of geothermal gradients for 20, 30, and 40 °C km⁻¹ from ref. 24 (solid black lines; red text labels) are also shown; intermediate contours (dashed lines) have been interpolated.

topographic effects by adding the difference between sample and mean local elevation (for a 10-km circle) to each closure depth determined by the method above. This effectively assumes constant topography since closure of the apatite He system. Erosion rates calculated in this way represent time-averaged rates since closure. Given the evidence for at least some surface uplift since the Late Miocene, it is therefore possible and even likely that many of these rates, especially those for samples with older ages, have varied. Nevertheless, assuming cooling results primarily from erosional exhumation, our technique captures the average rate at which samples transited the upper crust.

Calculated model erosion rates average 0.10 km Myr⁻¹ across the Washington Cascades at this latitude, but vary from as low as 0.02–0.04 km Myr⁻¹ on the eastern flank and near the topographic crest, to as high as 0.33 km Myr⁻¹ about two-thirds of the way up the west flank (Fig. 3). Calculated erosion rates for samples in some vertical transects display up to 0.08 km Myr⁻¹ differences between the highest and lowest elevation samples. This probably reflects faster erosion during the time intervals represented by these ages. This is supported by the slopes of the age–elevation relationships in

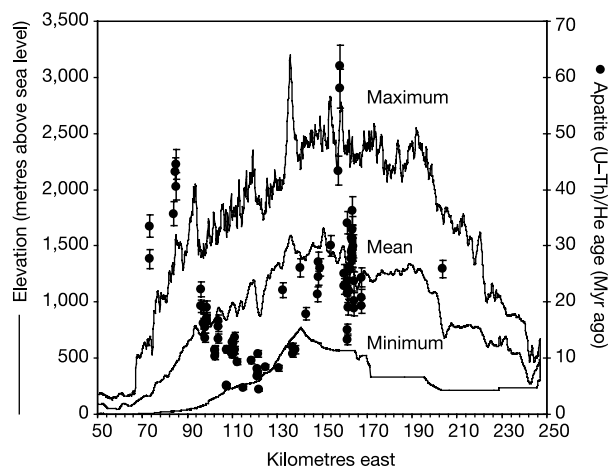


Figure 2 Apatite (U-Th)/He ages and topographic profiles (mean, maximum and minimum elevations) in swath across the Washington Cascades shown in Fig. 1. Ages younger than 10 Myr BP are restricted to the upper western flank of the range. Ages are oldest at the topographic crest or far flanks. Error bars are 6% (2 σ), based on reproducibility of standards. The isolated topographic high in the maximum topographic profile near 135 km east is Glacier Peak volcano.

some vertical transects, suggesting more rapid erosion (up to 0.7 km Myr⁻¹) from ~11–13 Myr BP. Despite these limitations to a steady erosion model, these data provide a useful estimate of long-term erosion rates, and more importantly, their differences across the range. If these rates have persisted since 10–15 Myr BP, approximately 3–5 km of rock have been removed from the mid-slope of the western flank, and only 0.5–1.0 km from other areas, including the topographic crest.

The absence of obvious active structures in the Washington Cascades that could produce spatially variable late Cenozoic rock uplift and erosion rates suggests that some other agent, such as climatic differences across the range, may be responsible for the order-of-magnitude difference in calculated erosion rates. It is reasonable to expect that one of the strongest climatic influences on local erosion rate would be the rate of mean annual precipitation (hereafter referred to simply as precipitation), through its effect on hillslope soil or regolith instability, or fluvial discharge and incision. Precipitation across the Washington Cascades shows a strong orographic effect, due to rising and cooling moist air masses encountering the topographic barrier of the range as they move eastward from the Pacific. Figure 3 shows the precipitation profile across the range centred at the latitude of Seattle, from interpolation of observed precipitation from 1961–1999 (ref. 24). In form and position relative to topography, the precipitation profile across the Cascades is remarkably similar to that of maximum calculated erosion rates (Fig. 3). Mean annual precipitation and calculated erosion reach a maximum of 3.5 myr⁻¹ and 0.33 km Myr⁻¹, respectively, at a mean elevation of about 1.1–1.2 km on the west flank of the range. Elsewhere in the range, including the topographic crest, precipitation is a factor of two to three lower, and calculated erosion rates are about three to fifteen times lower.

Local erosion rate might be expected to depend on several factors in addition to or instead of local precipitation, such as lithology, local slope or fluvial discharge. Bedrock across the entire transect is dominated by granitoid plutonic and gneissic rock, so lithologic variations would be unlikely to cause significant erosion rate variations. Likewise, mean local relief (as a proxy for local slope) in the transect shows little variation (1–1.3 km, for circles 10 km in diameter) across most of the range, and no significant correlation of relief with erosion rate is observed (Fig. 3). Fluvial discharge can be approximated to first order as integrated precipitation upstream,

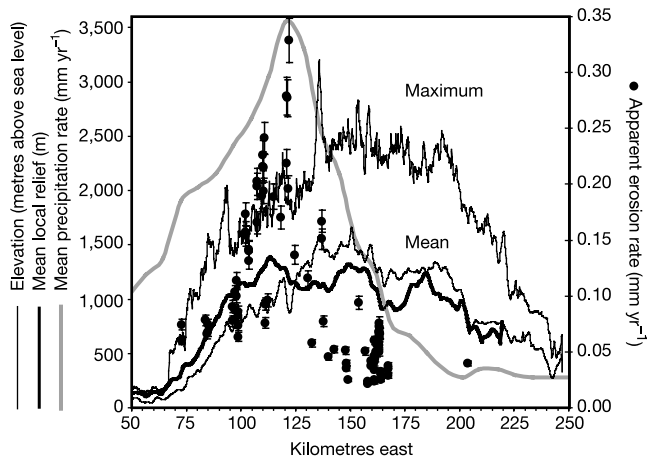


Figure 3 Comparison of topographic profiles, calculated model erosion rates, and mean annual precipitation rates. Thin black solid lines, mean and maximum topographic profiles. Filled symbols, erosion rates calculated from apatite (U–Th)/He ages. Bold grey line, profile of mean annual precipitation rate. Bold black line, mean local relief calculated for circles 10 km in diameter. Calculated erosion rates show a maximum of 0.33 km Myr^{-1} at mean elevations of about 1–1.2 km elevation on the west flank of the range. The mean local relief shows little variation over most of the range, but the mean annual precipitation rate largely follows mean erosion rate, with a maximum of 3.5 mm yr^{-1} in the same location as the erosion rate maximum. The precipitation rate maximum at ~120 km east is not an artefact of the orographic effect of Glacier Peak, because the precipitation transect is located near the latitude of Seattle and Mt Stuart, about 50 km to the south.

which is highest in the far western part of the transect, rather than at a mean elevation of ~1 km on the west flank, at the coincidence of erosion rate and precipitation maxima.

No major late Tertiary structures have been identified in the Cascades that could accommodate differential uplift and erosion rates across the range. It is possible that broad arching or folding could lead to higher uplift rates in the core of the range, but by itself (and in a steady state) this would be expected to lead to the highest erosion rates at the topographic crest, rather than 50 km to the windward side, coincident with the highest precipitation. Because of this, we conclude that the long-term erosion rate pattern across the range is controlled primarily by the precipitation pattern. The specific geomorphic processes coupling precipitation and erosion in this case are not clear.

Commonly used stream-power indices predict erosion as a function of main channel slope and discharge, but neither of these parameters correlate with precipitation or erosion rates inferred from apatite He ages. This may suggest a more important role for hill-slope processes or higher-order stream characteristics in controlling variations in long-term erosion rates, because these respond to more local variations in precipitation. Alternatively, variations in extents of glacial erosion across the range may contribute to the spatial pattern of erosion, because ice accumulation would also respond to local variations in precipitation. But, regardless of the specific geomorphic liaison between climate and erosion, these data suggest a predominately climatic influence on spatial variations in erosion rates across the range.

The strong variation in, and correlation between, precipitation and erosion rates across the Washington Cascades supports theoretical studies that argue for strong coupling and feedbacks between climate and tectonics in active orogens^{1–10}. Modelling studies suggest that mountain topography evolves towards a steady state, after which the macro-scale topography remains constant and erosion rates are equal to rock uplift rates. It is possible that rock

uplift and erosion rates are unrelated, in which case the modern Cascades topography is transient.

However, if the Cascades are in, or even close to, topographic steady state, then rock uplift on the west flank is as much as an order of magnitude faster than elsewhere in the range. The dynamic link by which rock uplift and deformation respond to spatially focused erosion could simply be isostasy, or some other mechanism such as accommodation of pluton emplacement or the vertical component of middle or lower crustal flow into crustal regions experiencing relatively rapid exhumation. In either case, the current orographic climate pattern is well correlated with, and may exert a strong influence on, the distribution of erosion, and possibly rock uplift and deformation, across the Cascades. □

Received 20 June; accepted 2 October 2003; doi:10.1038/nature02111.

1. Beaumont, C., Kooi, H. & Willett, S. in *Geomorphology and Global Tectonics* (ed. Summerfield, M. A.) 29–55 (John Wiley & Sons, Chichester, UK, 2000).
2. Willett, S. D. Orogeny and orography: The effects of erosion on the structure of mountain belts. *J. Geophys. Res.* **104**, 28957–28981 (1999).
3. Zeitler, P. K. *et al.* Erosion, Himalayan geodynamics, and the geomorphology of metamorphism. *GSA Today* **11**, 4–9 (2001).
4. Koons, P. O. The topographic evolution of collisional mountain belts: A numerical look at the southern Alps. *New Zealand. Am. J. Sci.* **289**, 1041–1069 (1989).
5. Montgomery, D. R. & Brandon, M. T. Topographic controls on erosion rates in tectonically active mountain ranges. *Earth Planet. Sci. Lett.* **201**, 481–489 (2002).
6. Beaumont, C., Fulsack, P. & Hamilton, J. in *Thrust Tectonics* (ed. McClay, K. R.) 1–18 (Chapman and Hall, New York, 1991).
7. Willett, S. D., Beaumont, C. & Fulsack, P. Mechanical model for the tectonics of doubly vergent compressional orogens. *Geology* **21**, 371–374 (1993).
8. Willett, S. D., Slingerland, R. & Hovius, N. Uplift, shortening, and steady state topography in active mountain belts. *Am. J. Sci.* **301**, 455–485 (2001).
9. Tucker, G. E. & Slingerland, R. L. Erosional dynamics, flexural isostasy, and long-lived escarpments: A numerical modeling study. *J. Geophys. Res.* **99**, 12229–12243 (1994).
10. Whipple, K. X., Kirby, E. & Brocklehurst, S. H. Geomorphic limits to climate-induced increases in topographic relief. *Nature* **401**, 39–43 (1999).
11. Brozovic, N., Burbank, D. W. & Meigs, A. J. Climatic limits on landscape development in the Northwestern Himalaya. *Science* **276**, 571–574 (1997).
12. Montgomery, D. R., Balco, G. & Willett, S. Climate, tectonics and the morphology of the Andes. *Geology* **29**, 579–582 (2001).
13. Smiley, C. J. The Ellensburg flora of Washington. *Univ. Cal. Publ. Geol. Sci.* **35**, 159–276 (1963).
14. Miller, M. M. *et al.* GPS-determination of along-strike variation in Cascadia margin kinematics: Implications for relative plate motion, subduction zone coupling, and permanent deformation. *Tectonics* **20**, 161–176 (2001).
15. Dragovich, J. D., *et al.* *Geologic Map of Washington—Northwest Quadrant*, Washington 3 sheets, 1:250,000, 71 pages (Geologic Map GM-50, Washington Division of Geology and Earth Resources, Washington State Department of Natural Resources), (2002).
16. Tabor, R. W. *et al.* *Geologic Map of the Skykomish River 30- by 60-minute Quadrangle*, Washington 1 sheet, 1:100,000, 42 pages (US Geol. Surv. Misc. Invest. Ser. Map I-1963, 1993).
17. Tabor, R. W., Frizzell, V. A. Jr, Booth, D. B. & Waitt, R. B. *Geologic Map of the Snoqualmie Pass 30 x 60 minute Quadrangle*, Washington 1 sheet, 1:100,000, 57 pages (US Geol. Surv. Misc. Invest. Ser. I-2538, 2000).
18. Tabor, R. W. *et al.* *Geologic Map of the Chelan 30-minute by 60-minute Quadrangle*, Washington 1 sheet, scale 1:100,000, 29 pages (US Geol. Surv. Misc. Invest. Ser. Map I-1661, 1987).
19. Tabor, R. W. *et al.* *Geologic Map of the Wenatchee 1:100,000 Quadrangle, Central Washington* 1 sheet, 1:100,000, 26 pages (US Geol. Surv. Misc. Invest. Ser. Map I-1311, 1982).
20. Farley, K. A. Helium diffusion from apatite: General behaviour as illustrated by Durango fluorapatite. *J. Geophys. Res.* **105**, 2903–2914 (2000).
21. Reiners, P. W. *et al.* Late Miocene exhumation and uplift of the Washington Cascades. *Geology* **30**, 767–770 (2002).
22. Brandon, M. T., Roden-Tice, M. K. & Garver, J. I. Late Cenozoic exhumation of the Cascadia accretionary wedge in the Olympic Mountains, northwest Washington State. *Geol. Soc. Am. Bull.* **110**, 985–1009 (1998).
23. Blackwell, D. D., Steele, J. L., Kelley, S. & Korosec, M. A. Heat flow in the state of Washington and thermal conditions in the Cascade range. *J. Geophys. Res.* **95**, 19495–19516 (1990).
24. Daly, C., Neilson, R. P. & Phillips, D. L. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteorol.* **33**, 140–158 (1994).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We acknowledge comments by P. Molnar. We also thank S. Willett and M. Brandon for discussions, and S. Nicolescu for analytical assistance. This work was supported by the US National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.W.R. (peter.reiners@yale.edu).

Links between erosion, runoff variability and seismicity in the Taiwan orogen

Simon J. Dadson¹, Niels Hovius¹, Hongey Chen², W. Brian Dade⁴, Meng-Long Hsieh², Sean D. Willett⁵, Jyr-Ching Hu², Ming-Jame Horng⁶, Meng-Chiang Chen⁷, Colin P. Stark⁸, Dimitri Lague¹ & Jiun-Chuan Lin³

¹Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge, CB2 3EQ, UK

²Department of Geosciences, ³Department of Geography, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei, Taiwan

⁴Department of Earth Sciences, Dartmouth College, Hanover, New Hampshire 03755, USA

⁵Department of Earth and Space Sciences, University of Washington, Seattle, Washington 98195, USA

⁶Water Resources Agency, Ministry of Economic Affairs, Hsin-Yi Road, Taipei, Taiwan

⁷Taroko National Park Headquarters, Fu-Su Village, Hualien, 972, Taiwan

⁸Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York 10964, USA

The erosion of mountain belts controls their topographic and structural evolution^{1–3} and is the main source of sediment delivered to the oceans⁴. Mountain erosion rates have been estimated from current relief and precipitation, but a more complete evaluation of the controls on erosion rates requires detailed measurements across a range of timescales. Here we report erosion rates in the Taiwan mountains estimated from modern river sediment loads, Holocene river incision and thermochronometry on a million-year scale. Estimated erosion rates within the actively deforming mountains are high (3–6 mm yr^{–1}) on all timescales, but the pattern of erosion has changed over time in response to the migration of localized tectonic deformation. Modern, decadal-scale erosion rates correlate with historical seismicity and storm-driven runoff variability. The highest erosion rates are found where rapid deformation, high storm frequency and weak substrates coincide, despite low topographic relief.

Taiwan results from collision of the Luzon arc, on the Philippine sea plate, and the Asian continental margin⁵ (Fig. 1). The 4-km-high mountain belt links opposite-dipping subduction systems at the Manila and Ryukyu trenches. Plate boundary obliquity has led the collision to propagate southward over the past 5 Myr, with marine terrace uplift⁶ and exhumation rates⁷ of 5–7 mm yr^{–1}. Across central Taiwan metamorphic grade increases from poorly consolidated, Upper Tertiary sediments in the Western Foothills thrust belt, through slates in the Hsuehshan and western Central ranges, to greenschist grade pre-Tertiary meta-sediments in the eastern Central Range. The subtropical climate of Taiwan, with an average of four typhoons per year⁸ and mean annual precipitation of 2.5 myr^{–1}, combined with frequent earthquakes, together drive rapid mass-wasting and fluvial bedrock incision^{9,10}.

Since 1970, river suspended-sediment discharge has been recorded at over 150 stations across Taiwan¹¹. We have estimated average catchment-wide erosion rates^{12,13} from these measurements for the interval 1970–1999 (Methods; Supplementary Fig. S1a; Supplementary Table S1). Erosion rates estimated from repeat bathymetric surveys of water reservoirs (1970–1998) correlate with suspended-load estimates in nearby catchments ($r^2 = 0.45$; Supplementary Fig. S1b), indicating that suspended sediment and bedload comprise about $70 \pm 28\%$ and $30 \pm 28\%$, respectively, of total river load in high mountains (uncertainties are 95% confidence intervals).

The 30-yr average suspended-sediment erosion rate of Taiwan is

3.9 mm yr^{–1}. Between 1970 and 1999, Taiwan supplied 384 Mt yr^{–1} of suspended sediment to the ocean. This represents 1.9% of estimated global suspended sediment discharge⁴ and yet is derived from only 0.024% of Earth's subaerial surface. Adding 30% bedload would increase the estimated erosion rate to 5.2 mm yr^{–1} and average annual sediment yield to 500 Mt yr^{–1}, although much of the bedload is trapped in floodplains before reaching the sea.

Decadal-scale erosion rates were high in the eastern Central Range and southwest Taiwan, but low in the north and west (Fig. 2a). The metamorphic core is eroding at ~ 6 mm yr^{–1}. Erosion rates of up to 60 mm yr^{–1} were found around active thrust faults in the south of the Western Foothills thrust belt. North and west Taiwan had lower erosion rates (1–4 mm yr^{–1}).

To complement these modern, catchment-scale erosion rates, we obtained reach-scale river incision rates from 25 Holocene strath terraces at 12 locations, dated using ¹⁴C (Methods; Supplementary Table S2; Fig. 2b). Channel incision rates range from 1.5 mm yr^{–1} in the north and 9 mm yr^{–1} in the east, to over 15 mm yr^{–1} around active structures in the Western Foothills thrust belt¹⁴.

Persistence of erosion rates over million-year timescales was assessed using new and published¹⁵ apatite fission-track data with a geothermal model to calculate exhumation rates (Methods; Fig. 2c). Samples whose fission-track age was reset in the present orogeny (≤ 5 Myr) were found in the eastern section of the thrust belt and the metamorphic core of the mountain belt. They indicate exhumation rates of 3–6 mm yr^{–1} in the eastern Central Range

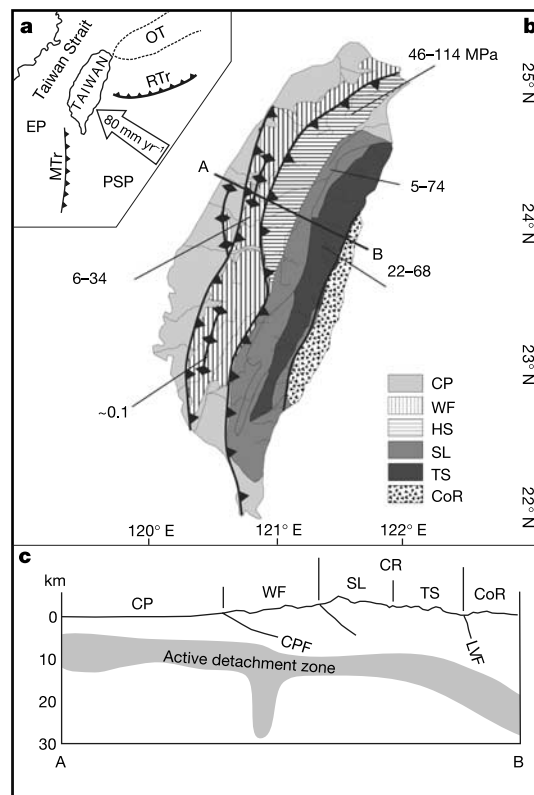


Figure 1 Location and geotectonic framework of Taiwan. **a**, Plate tectonics of the Ryukyu–Taiwan–Luzon area⁵. EP, Eurasian plate; PSP, Philippine Sea plate; MTr, Manila trench; RTr, Ryukyu trench; OT, Okinawa trough. **b**, Lithology²⁸ with structures and uniaxial compressive rock strength measurements given in MPa. CP, coastal plain; WF, Western Foothills thrust belt; HS, Hsuehshan Range; SL, slate belt; TS, Tananao schist; CoR, Coastal Range. **c**, Schematic cross-section⁶ showing basal detachment zone imaged using small earthquakes²⁹. CR, Central Range; CFF, Chelungpu fault; LVF, longitudinal valley fault.

and $1.5\text{--}2.5\text{ mm yr}^{-1}$ in the eastern thrust belt and south-central Taiwan. In contrast, unreset apatite ages from the Western Foothills thrust belt and south Taiwan reflect earlier exhumation ($>5\text{ Myr}$). No rate information can be obtained from unreset samples.

Taken together, our data indicate that erosion rates in the eastern Central Range were $3\text{--}6\text{ mm yr}^{-1}$ across all timescales considered here, although some catchments had higher decadal erosion rates. This persistent, rapid erosion has exhumed greenschist-grade metamorphic rocks in east Taiwan. By comparison, rapid Holocene

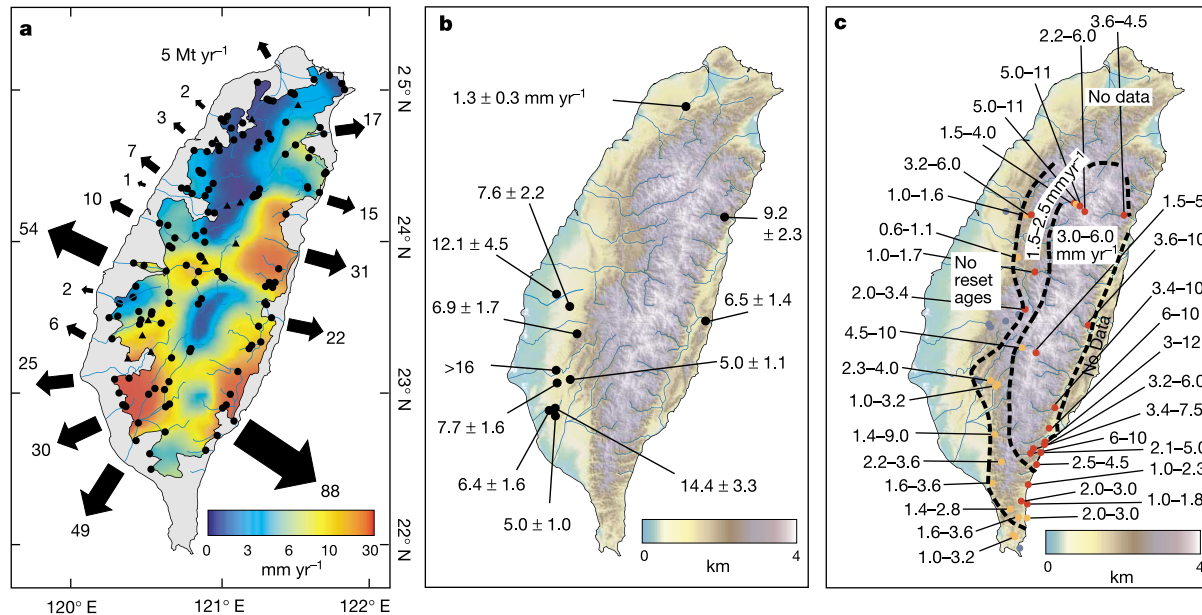


Figure 2 Erosion rates in Taiwan across multiple timescales. **a**, Calculated from fluvial suspended sediment observations with 5-km grid resolution, smoothed at catchment scale using a circular moving mean with 30-km diameter. Black arrows indicate mean annual coastal suspended sediment flux from rivers draining areas greater than 400 km^2 . Coastal sediment fluxes are reported from hydrometric stations closest to coast. Average uncertainty of these values is 37%. Black circles indicate hydrometric stations used to

construct the map; triangles indicate locations of reservoirs used in reservoir fill analysis; no data are available within the grey shaded area. **b**, Bedrock strath incision rates (all in mm yr^{-1}). Values for each locality represent mean incision rate for all terraces measured at that locality. Uncertainties are one standard deviation in radiometric age and 20% in strath elevation. **c**, Exhumation rates (all in mm yr^{-1}) calculated from apatite fission-track ages: red, reset; orange, partially reset; blue, unreset.

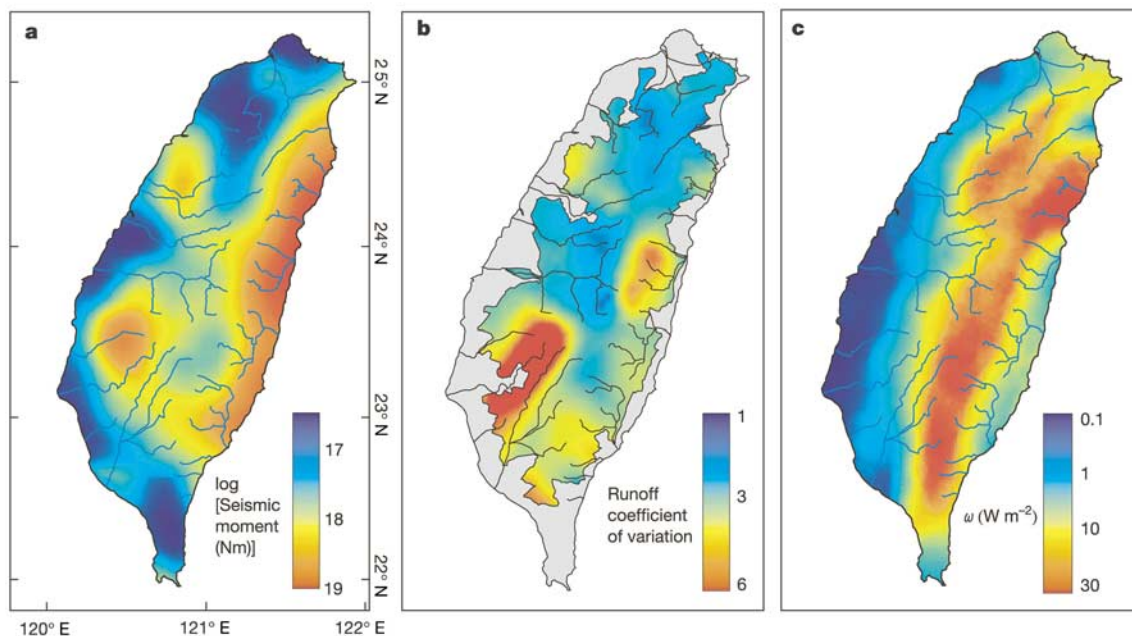


Figure 3 Seismic, hydrological and topographic controls on denudation pattern. **a**, Cumulative seismic moment, M_0 , computed from historical record of earthquakes greater than $M_w = 5.0$, between 1900 and 1998 (that is, excluding the 1999 Chi-Chi earthquake). Where seismic moment is not reported directly, we have estimated it from the listed magnitude using a global relation³⁰. **b**, Runoff coefficient of variation

(dimensionless), defined as standard deviation of runoff divided by mean runoff. Runoff (m yr^{-1}) was measured as average annual river discharge divided by drainage area. **c**, Unit stream-power (W m^{-2}) presented at 1-km grid resolution with 30-km circular moving mean applied.

river incision rates of 5–12 mm yr⁻¹, and decadal erosion rates averaging 15 mm yr⁻¹, are probably a recent feature associated with westward propagation of the thrust belt. Unreset apatite fission-track ages in the Western Foothills thrust belt appear to confirm this view, although thermochronometers are relatively insensitive to the shallow, largely horizontal deformation of the thrust belt. Rapid fluvial incision in the Western Foothills correlates with active, blind thrusts and fault-ramp systems in poorly consolidated Pliocene–Quaternary sediments^{16,17}, and we hypothesize that high catchment-wide erosion rates have suppressed the topographic expression of these faults.

Patterns of Holocene incision rate and decadal suspended-sediment erosion rate are broadly consistent, but with maximum values of 60 mm yr⁻¹, decadal rates systematically exceed Holocene incision rates in southwest Taiwan. Moreover, decadal erosion rates are lower than expected in the north of the Western Foothills thrust belt. Explanations include changing climate and land-use, or brevity of the hydrometric record. Palynology indicates substantial climate variation¹⁸ across Taiwan during the Holocene, but this is unlikely to account for the north–south contrast. Growth of agriculture and urbanization may have affected erosion, although urbanization is limited to the coastal plains. Our data show no trend of increased erosion with economic development. We propose, instead, that the decadal erosion pattern reflects the location of recent climatic and seismic perturbations, superimposed on other, persistent controls on erosion.

We have evaluated controls on erosion rates in Taiwan using topographic, climatic, geomechanical and seismic databases. Spatial variability in average erosion rates can be explained to a statistically significant degree (with 95% confidence) by (1) cumulative seismic moment release ($r^2 = 0.13$; Fig. 3a) and (2) temporal variability in runoff ($r^2 = 0.27$; Fig. 3b). The bulk of the remaining variation probably originates in the natural variability in mountain drainage basins. Surprisingly, no significant correlation was found between decadal erosion rates and stream power ($r^2 = 0.01$; Fig. 3c), precipitation ($r^2 = 0.001$), runoff ($r^2 = 0.03$) or local relief ($r^2 = 0.01$) (Supplementary Fig. S2). These results indicate that modern erosion rates are strongly influenced by large earthquakes and typhoons, both of which drive widespread sediment supply to channels.

Earthquakes produce sediment by rock mass shattering and landsliding¹⁹. In Taiwan there is a clear link between erosion and historical cumulative seismic moment release between 1900 and 1998 (Fig. 4a). Historical seismicity was high along the east coast and across southern Taiwan; but low in the north and west, in the far south, and in a small region of the Central Range (Fig. 3a). This pattern agrees with GPS convergence rates^{16,20}. Across Taiwan, cumulative seismic moment release correlates linearly with decadal erosion rate over a sixfold range. Seismic sediment production is therefore an important control on erosion in Taiwan. We note, however, that the seismically driven erosion pattern may be relatively short-lived, given the along-strike uniformity of shortening across the mountain belt. The southwestern section of the thrust belt experienced eleven historic earthquakes of magnitude $M_w > 6.0$ between 1900 and 1998, whereas only three such earthquakes occurred in the north. The 1999 $M_w = 7.6$ Chi-Chi earthquake may be expected to elevate erosion rates considerably in the north of Taiwan.

Sediment is also produced and mobilized by storm-triggered landslides. This is clear from the significant correlation between erosion rates and the coefficient of runoff variation, which represents storm incidence (Fig. 4b). Runoff variability is high along the east coast of Taiwan, where most incoming typhoons make landfall; it is especially high in the south of Taiwan, where typhoon passage is not obstructed by topography (Fig. 3b). We conclude that storm runoff is a first-order control on erosion rates in Taiwan, triggering primary landslides and debris flows, which flush colluvial

sediment from the mountain belt.

A common assumption is that erosion rates are jointly governed by relief and average precipitation^{21,22}. This assumption does not fit modern erosion rates in Taiwan. We have compared the observed decadal erosion pattern with stream power, a combined measure of relief and precipitation (Methods). The stream power pattern (Fig. 3c) emphasizes areas of high relief and orographic precipitation along the Central Range, where rapid exhumation has persisted longest. However, it does not match first-order features of the decadal erosion pattern, such as high erosion rates at mountain fronts, the order-of-magnitude increase from north to south, and high erosion rates around active thrusts in the southwest. We hypothesize that, over decadal timescales, stream power is not the rate-limiting control on erosion in Taiwan. This is due to intermittent storm- and earthquake-driven sediment supply, and the effect of substrate strength on topographic relief.

The effect of substrate strength was assessed using 1,114 measurements of uniaxial compressive strength at 23 sites across Taiwan (Fig. 1b; Supplementary Table S3). Rock strength varies over three orders of magnitude (0.1–253 MPa), broadly increasing with metamorphic grade from west to east, and south to north, with a drop in the alluviated depression between the Central and Coastal ranges. This trend is reflected in the relief of Taiwan, which is greatest in competent meta-sandstones and steepest in marbles and gneisses in

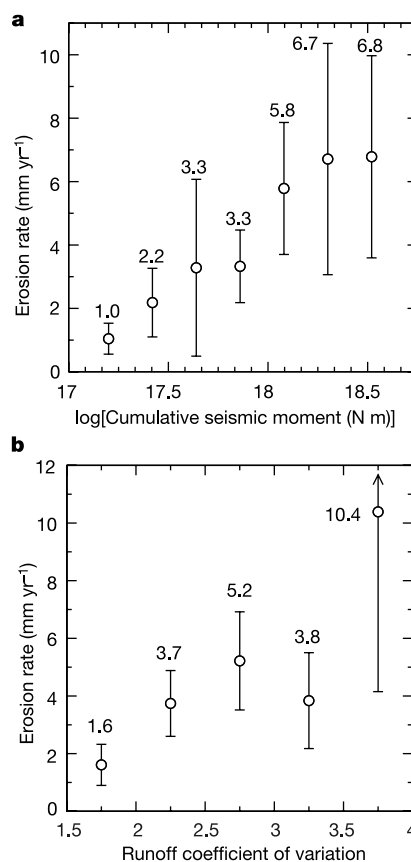


Figure 4 Correlations between erosion rates and their controls. **a**, Drainage-basin-wide suspended-sediment erosion rate as a function of cumulative scalar seismic moment, M_0 . In total 128 drainage basins are shown (two extreme outliers were removed), binned by mean cumulative seismic moment release within their boundaries. **b**, Drainage-basin-wide suspended-sediment erosion rate as a function of coefficient of runoff variation. In total, 128 drainage basins are shown (two extreme outliers were removed), binned by the coefficient of runoff variation defined by their water discharge record. For **a** and **b**, the mean erosion rates within these bins are shown, and their 95% confidence intervals.

the northern Central Range, but subdued in Pliocene–Quaternary sediments of the Western Foothills thrust belt. Across Taiwan, compressive strength correlates with average slope ($r^2 = 0.31$) and basin relief ($r^2 = 0.17$). The highest erosion rates are in low-relief terrain with weak substrate. This reinforces the notion that substrate properties modulate the topographic expression of tectonic deformation and erosion. Determining erosion rates from topography is possible only when spatially variable substrate erodibility is taken into account.

In summary, progressive exhumation of competent rocks at rates up to $3\text{--}6\text{ mm yr}^{-1}$ has built a high and steep mountain belt in Taiwan. Holocene river incision rates are highest on active faults. Erosion modifies the topographic expression of active faults to a degree determined by the strength of the exposed substrate. The decadal erosion pattern is strongly associated with production of sediment by landsliding during large earthquakes and storms. Modern erosion follows changing patterns of seismic deformation, suggesting that future erosion rates in north and central Taiwan will be elevated following the 1999 $M_w = 7.6$ Chi-Chi earthquake. □

Methods

River sediment load

At each hydrometric station, water discharge was measured daily and suspended sediment concentration was measured at an average frequency of 24 samples per year using a USDH-48 depth-integrating suspended sediment sampler²³ (Supplementary Fig. S1a). We used a bias-corrected average of the reported measurements to obtain time-averaged suspended sediment loads²⁴. This method corrects for the increased number of observations in the summer months. An alternative approach using a rating curve between water and suspended sediment discharge is not appropriate for most rivers in Taiwan because it assumes that rivers are transport-limited. In Taiwan, most catchments have a strong supply-limited component⁹, so a direct average as used here is the most appropriate method for estimating suspended sediment discharge²⁴. To obtain a stable measure of erosion we used only observations from the 130 gauging stations with records spanning at least eight years. Catchment-wide erosion rates were calculated from annual suspended sediment discharge by dividing by the density of quartz and by drainage area. For catchments that had several nested gauges we recalculated erosion rates for the parts with multiple gauges, using data from each upstream station. Our analysis extends until the end of August 1999, excluding the effects of the Chi-Chi earthquake. At present the onward transport of material produced during the Chi-Chi earthquake is incomplete. Extension of our analysis to the end of 2001 would lead to a 12% increase in our estimate of the annual average suspended sediment discharge from Taiwan. We do not consider chemical erosion rates, which are probably an order of magnitude lower than physical erosion rates¹² and are poorly constrained.

Channel incision rates

Bedrock incision rates were obtained by dating organic material deposited on strath terraces formed as rivers incised through their bedrock substrate. Measurements are from bedrock terraces overlain by less than five metres of channel deposit. Radiometric ages were calculated from ^{14}C in wood or plant fragments in the alluvial veneer, and converted to calendar ages²⁵. Incision rates were calculated by dividing the elevation of the bedrock terrace above the modern river by the calendar age. Uncertainties were calculated as 1σ for age and as 20% for terrace elevation, the latter to account for variations in river stage.

Fission track ages and exhumation rates

Long-term exhumation rates were obtained from apatite fission track ages. Fission-induced damage tracks in apatite are annealed at million-year timescales at temperatures over 110°C , so that fission track ages reflect cooling associated with exhumation from depths of 3–5 km. The exhumation rates in Fig. 2c were determined using published ages from the Central Range¹⁵ and unpublished ages from the Western Foothills thrust belt, interpreted through a thermal model with one-dimensional heat transfer, an initial geothermal gradient of 25°C km^{-1} , a surface temperature of 10°C and a constant erosion rate. Lacking an independent estimate of the duration of exhumation, we use the first occurrence of reset ages at the surface. This assumption provides a maximum estimate of exhumation rate. If geothermal gradients are higher owing to a greater duration of exhumation or conductive focusing of heat into valleys where samples were primarily obtained, the inferred exhumation rates would be lower.

Stream power erosion index

Stream power per unit area of bed, ω (W m^{-2}), can be written^{26,27} as $\omega = \rho g Q S / w$, where ρ is water density ($1,000\text{ kg m}^{-3}$), g is acceleration due to gravity (9.8 m s^{-2}), Q is water discharge ($\text{m}^3\text{ s}^{-1}$), S is channel slope, and w is channel width (m). An index of erosion potential based on stream power can be constructed using a combination of river discharge data and attributes derived from a digital elevation model (DEM)^{21,22}. We computed unit stream power using a 40-m resolution DEM to measure slope in the direction of steepest descent. Runoff was estimated from water discharge measurements and annual discharge was calculated for each point in the DEM by integrating runoff upslope. Channel width cannot be resolved using a DEM of this scale, so we used its scaling

relation with discharge ($w \propto Q^{0.5}$) to rewrite unit stream power as $\omega = \rho g Q^{0.5} S$. This is confirmed for Taiwan by measurements of channel width from a 5-m resolution satellite image at 12 gauging stations in central Taiwan.

Received 23 May; accepted 21 October 2003; doi:10.1038/nature02150.

- Koons, P. O. The topographic evolution of collisional mountain belts: A numerical look at the Southern Alps, New Zealand. *Am. J. Sci.* **289**, 1041–1069 (1989).
- Beaumont, C., Fullsack, P. & Hamilton, J. In *Thrust Tectonics* (ed. McClay, K. R.) 1–18 (Chapman and Hall, London, 1992).
- Willett, S. D. Orogeny and orography: the effects of erosion on the structure of mountain belts. *J. Geophys. Res.* **104**, 28957–28981 (1999).
- Milliman, J. D. & Syvitsky, J. P. M. Geomorphic/tectonic control of sediment discharge to the ocean: The importance of small mountainous rivers. *J. Geol.* **100**, 525–544 (1992).
- Teng, L. S. Geotectonic evolution of late Cenozoic arc-continent collision in Taiwan. *Tectonophysics* **183**, 57–76 (1990).
- Liew, P. M. et al. Holocene tectonic uplift deduced from elevated shorelines, eastern Coast Range of Taiwan. *Tectonophysics* **222**, 55–68 (1993).
- Liu, T. K. Tectonic implications of fission-track ages from the Central Range, Taiwan. *Proc. Geol. Soc. China* **25**, 22–37 (1982).
- Shieh, S. L. et al. *Users' Guide for Typhoon Forecasting in the Taiwan Area (VIII)* (Central Weather Bureau, Taipei, 2000).
- Hovius, N., Stark, C. P., Chu, H. T. & Lin, J. C. Supply and removal of sediment in a landslide-dominated mountain belt: Central Range, Taiwan. *J. Geol.* **108**, 73–89 (2000).
- Hartshorn, K., Hovius, N., Dade, W. B. & Slingerland, R. L. Climate-driven bedrock incision in an active mountain belt. *Science* **297**, 2036–2038 (2002).
- Water Resources Agency. *Hydrological Yearbook of Taiwan, Republic of China* (Ministry of Economic Affairs, Taipei, 1970–2001); at <http://gweb.wra.gov.tw/wrweb/>.
- Li, Y. H. Denudation of Taiwan island since the Pleistocene epoch. *Geology* **4**, 105–107 (1976).
- Fuller, C. W., Willett, S. D., Hovius, N. & Slingerland, R. Erosion rates for Taiwan mountain basins: New determinations from suspended sediment records and a stochastic model of their temporal variation. *J. Geol.* **111**, 71–87 (2003).
- Hsieh, M. L. & Knuepfer, P. L. K. Middle-late Holocene river terraces in Erjen River basin, southwestern Taiwan—implications of river response to climate change and active tectonic uplift. *Geomorphology* **38**, 337–372 (2001).
- Willett, S. D., Fisher, D., Fuller, C., Yeh, E. C. & Lu, C. Y. Erosion rates and orogenic wedge kinematics in Taiwan inferred from apatite fission track thermochronometry. *Geology* **31**, 945–948 (2003).
- Hickman, J. B., Wiltchko, D. V., Hung, J. H., Fang, P. & Bock, Y. In *Geology and Geophysics of an Arc-Continent Collision, Taiwan* (eds Byrne, T. B. & Liu, C. S.) 75–92 (Geological Society of America Special Paper 358, GSA, Boulder, CO, 2002).
- Hsieh, M. L. & Knuepfer, P. L. K. In *Geology and Geophysics of an Arc-Continent Collision, Taiwan* (eds Byrne, T. B. & Liu, C. S.) 55–74 (Geological Society of America Special Paper 358, GSA, Boulder, CO, 2002).
- Liew, P. M., Kuo, C. M., Huang, S. Y. & Tseng, M. H. Vegetation change and terrestrial carbon storage in eastern Asia during the Last Glacial Maximum as indicated by a new pollen record from central Taiwan. *Glob. Planet. Change* **16–17**, 85–94 (1998).
- Keefer, D. K. Landslides caused by earthquakes. *Geol. Soc. Am. Bull.* **95**, 406–421 (1984).
- Yu, S. B., Chen, H. Y. & Kuo, L. C. Velocity field of GPS stations in the Taiwan area. *Tectonophysics* **274**, 41–59 (1997).
- Montgomery, D. R., Balco, G. & Willett, S. D. Climate, tectonics, and the morphology of the Andes. *Geology* **27**, 579–582 (2001).
- Finlayson, D. P., Montgomery, D. R. & Hallet, B. Spatial coincidence of rapid inferred erosion with young metamorphic massifs in the Himalayas. *Geology* **30**, 219–222 (2002).
- Kao, S. J. & Liu, K. K. Estimating the suspended sediment load by using the historical hydrometric record from the Lanyang-Hsi watershed. *Terrest. Atmos. Oceanic Sci.* **12**, 401–414 (2001).
- Cohn, T. A. Recent advances in statistical methods for the estimation of sediment and nutrient transport in rivers. *Rev. Geophys.* **33** (suppl.) 1–18 (1995).
- Stuiver, M. & Raimer, P. J. Extended ^{14}C data base and revised CALIB 3.0 ^{14}C age calibration program. *Radiocarbon* **35**, 215–230 (1993).
- Bagdold, R. A. The flow of cohesionless grains in fluids. *Phil. Trans. R. Soc. Lond. A* **249**, 335–397 (1956).
- Whipple, K. X. & Tucker, G. E. Dynamics of the stream-power river incision model: Implications for height limits of mountain ranges, landscape response timescales, and research needs. *J. Geophys. Res.* **104**, 17661–17674 (1999).
- Ho, C. S. *An Introduction to the Geology of Taiwan: Explanatory Text of the Geological Map of Taiwan* 2nd edn (Central Geological Survey, Ministry of Economic Affairs, Taipei, 1986).
- Carena, S., Suppe, J. & Kao, H. Active detachment of Taiwan illuminated by small earthquakes and its control of first-order topography. *Geology* **30**, 935–938 (2002).
- Ekström, G. & Dziewonski, A. M. Evidence of bias in the estimation of earthquake size. *Nature* **332**, 319–323 (1988).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by the UK Natural Environment Research Council with a CASE addition from Faber Maunsell plc, and by the Taiwan National Science Council, and the Royal Society. We thank P. Molnar and D. Burbank for constructive reviews, and N. Lang and S. Pegg for assistance with data analysis.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.J.D. (simon00@esc.cam.ac.uk).

Decoupling of erosion and precipitation in the Himalayas

D. W. Burbank¹, A. E. Blythe², J. Putkonen³, B. Pratt-Sitaula¹, E. Gabet¹, M. Oskin¹, A. Barros⁴ & T. P. Ojha⁵

¹Department of Geological Sciences, University of California, Santa Barbara, California 93106, USA

²Department of Earth Sciences, University of Southern California, Los Angeles, California 90089, USA

³Department of Earth and Space Sciences, University of Washington, Seattle, Washington 98195, USA

⁴Division of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA

⁵Himalayan Experience, 659-1 Mhepi, Kathmandu 3, Nepal

The hypothesis that abrupt spatial gradients in erosion can cause high strain rates in active orogens has been supported by numerical models that couple erosional processes with lithospheric deformation via gravitational feedbacks^{1–3}. Most such models invoke a ‘stream-power’ rule, in which either increased discharge or steeper channel slopes cause higher erosion rates. Spatial variations in precipitation and slopes are therefore predicted to correlate with gradients in both erosion rates and crustal strain. Here we combine observations from a meteorological network across the Greater Himalaya, Nepal, along with estimates of erosion rates at geologic timescales (greater than 100,000 yr) from low-temperature thermochronometry. Across a zone of about 20 km length spanning the Himalayan crest and encompassing a more than fivefold difference in monsoon precipitation, significant spatial variations in geologic erosion rates are not detectable. Decreased rainfall is not balanced by steeper channels. Instead, additional factors that influence river incision rates, such as channel width and sediment concentrations, must

compensate for decreasing precipitation. Overall, spatially constant erosion is a response to uniform, upward tectonic transport of Greater Himalayan rock above a crustal ramp.

Contrasting styles of strain on the southern versus the northern flanks of the Tibetan plateau suggest potential responses to precipitation variations: within the narrow Himalayan belt, monsoon precipitation has removed mass and localized deformation since Miocene times⁴, whereas ineffective erosion under drier climates has led to accretion and northward migration of deformation in northern Tibet⁵. Such interpretations of tectonic–precipitation linkages, as well as predictions of coupled geodynamic–erosion models, can only be tested against field data in situations where the measurable signals rise well above the noise of natural systems. The Nepalese Himalaya appear well suited for such a test, because they encompass abrupt gradients in precipitation, erosion and deformation.

The Marsyandi River catchment in the Annapurna Himalaya (Fig. 1) is one of the few drainages that cuts nearly orthogonal to Himalayan structural trends, and extends without major deviations from the southern edge of the Tibetan plateau to the Himalayan foreland. From its headwaters in Tibetan sedimentary strata, the Marsyandi crosses the South Tibetan fault (STF) system^{4,6} before traversing the Greater Himalayan crystalline belt in a gorge incised between peaks rising to >7,000 m (Fig. 2a). On crossing the Main Central thrust (MCT), the Marsyandi flows through the Lesser Himalaya and crosses the Main Boundary Thrust (MBT) as it enters the Ganges basin. Both the STF system, a primarily down-to-the-north normal fault system, and the MCT, an up-to-the-south thrust fault, experienced major displacement in the early Miocene (~20–15 Myr ago)^{4,6}. The MCT zone comprises both the classic MCT (‘MCT I’, Fig. 1), as well as a swath ~10 km wide (‘MCT zone’, Fig. 1) in the northernmost part of the MCT I footwall, where young cooling ages indicate accelerated Plio-Pleistocene erosion^{7,8}. The likelihood of Quaternary displacement in the MCT zone and the STF system is currently debated^{4,9,10}.

A newly established meteorological network (Fig. 1) spans from

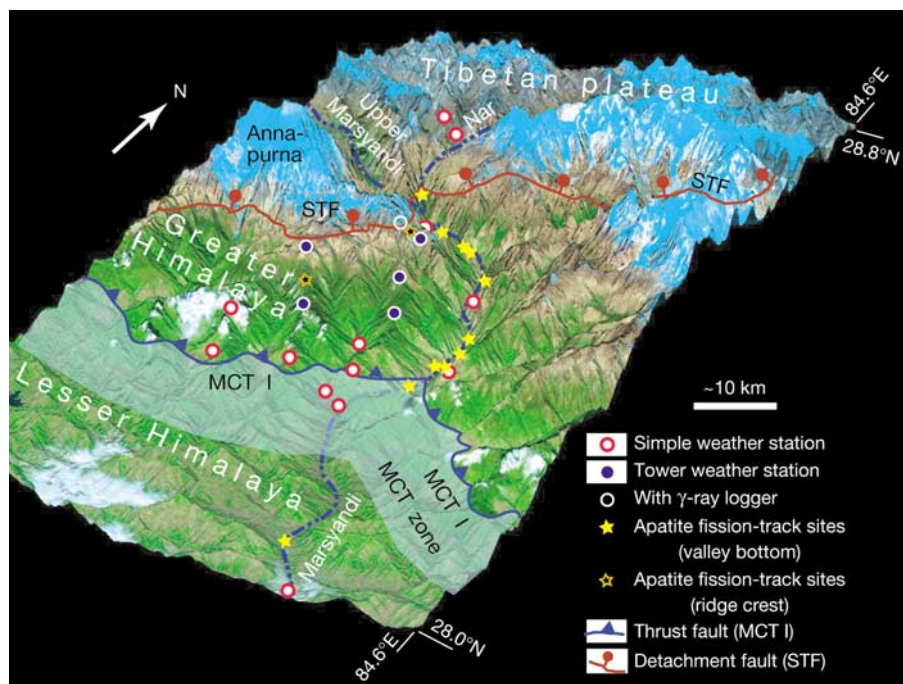


Figure 1 Three-dimensional perspective view of study area, showing locations of major geologic structures, meteorological stations, and apatite fission-track samples. Sites with 10-m-tall weather towers or with γ -ray loggers are distinguished from ‘simple’

stations. Approximate traces of the MCT I and STF system are shown, along with the ‘MCT zone’ delineating the northern footwall of the MCT I.

Table 1 Apatite fission-track ages in the Marsyandi catchment

Sample	Topographic position	Age (Myr)	1 σ (Myr)	Latitude, N	Longitude, E	Elevation
NBE-2	Valley	3.8	0.5	28° 01.473'	84° 27.500'	474
NBE-4	Valley	1.9	1.1	28° 19.538'	84° 24.095'	900
FT-JP-00-07	Valley	0.7	0.2	28° 23.242'	84° 24.173'	1,210
NBE-5	Valley	0.4	0.2	28° 23.482'	84° 24.311'	1,100
NBE-13	Valley	0.3	0.2	28° 25.334'	84° 24.115'	1,320
NBE-11	Valley	0.5	0.2	28° 30.637'	84° 21.641'	1,850
NBE-9	Valley	0.9	0.3	28° 31.763'	84° 20.615'	2,280
FT-4-97	Valley	0.0	0.2	28° 31.800'	84° 19.253'	2,300
NBE-10	Valley	0.5	0.2	28° 31.997'	84° 20.615'	2,150
NBE-8a	Valley	0.4	0.2	28° 33.683'	84° 15.640'	2,640
FT-1-97	Ridge crest	1.2	0.3	28° 29.616'	84° 18.205'	4,100
JP-FT-01-2	Ridge crest	0.8	0.2	28° 22.971'	84° 15.236'	3,831

500 to 1,500 m in the Lesser Himalaya, rises to 4,500 m in the Greater Himalaya, and includes one site at 4,200 m in the Tibetan zone. The 4-yr record (Fig. 2b) indicates that precipitation in the Annapurna Himalaya is dominated by monsoon rainfall: 80–98% of the annual total accumulates between May and October. Rainfall in the Lesser Himalaya averages $\sim 1.6 \text{ m yr}^{-1}$, but abruptly increases as the monsoon impinges on the southern Greater Himalayan slopes, where maximum monsoon rainfall ($\sim 4.3 \text{ m}$) occurs at $\sim 3,000 \text{ m}$ elevation. Striking variations occur perpendicular to the range as monsoon precipitation systematically decreases northward to $<0.5 \text{ m}$ in the Tibetan realm (Fig. 2b).

Spatial variations in erosion rates across the Himalaya at geologic timescales were assessed using apatite fission-track dating on bed-rock samples spanning 50 km along the Marsyandi River from the southern edge of the Tibetan zone to the northern margin of the Lesser Himalaya (Table 1). Apatite fission-track ages are commonly interpreted to indicate the elapsed time since cooling below the annealing isotherm. When cooling is rapid ($\geq 100^\circ \text{C Myr}^{-1}$), as is usual here, the effective annealing temperature is $\sim 140^\circ \text{C}$ (ref. 11). If the depth of the annealing isotherm and the trajectory of rocks towards the surface are estimated, apatite fission-track ages can be converted into mean erosion rates^{12,13}. Even without a three-dimensional thermal and kinematic model, cooling rates obtained from analogous topographic positions (river bottoms) can be compared as proxies for relative erosion rates.

The apatite fission-track dates show a striking pattern in the context of Himalayan topography and rainfall (Fig. 2c). At the southernmost site, a $\sim 4\text{-Myr}$ age yields a cooling rate of $\sim 30^\circ \text{C Myr}^{-1}$ for the northern part of the Lesser Himalaya. Apatite fission-track ages become younger and cooling rates more rapid as the Greater Himalaya are approached. In the MCT hanging wall, apatite fission-track ages abruptly drop to $\sim 0.5 \text{ Myr}$, indicating cooling rates nearly an order of magnitude faster than in the nearby Lesser Himalaya. Across the $\sim 20 \text{ km}$ separating the MCT from the STF, the apatite fission-track ages show no systematic trend, but instead fluctuate around a mean of $\sim 0.5 \pm 0.2 \text{ Myr}$. These data support a key conclusion: cooling rates at million-year timescales

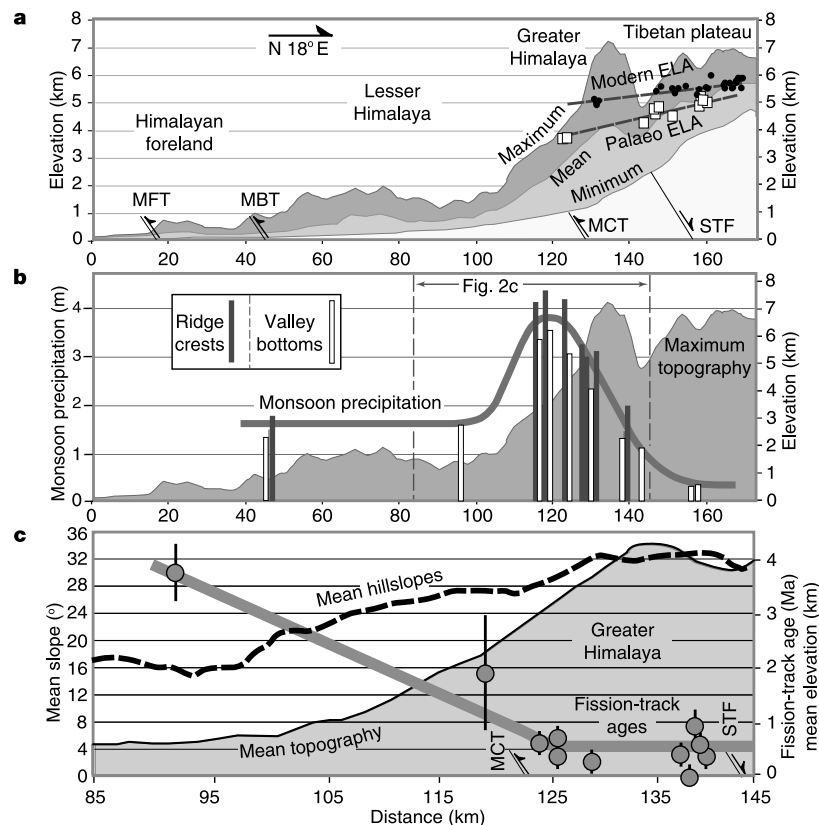


Figure 2 Monsoon precipitation, apatite fission-track ages, glacial equilibrium-line altitude (ELA) and topographic characteristics of the Marsyandi drainage. MFT, Main Frontal thrust. **a**, Maximum, minimum, and mean elevation within a 50-km-wide swath oriented parallel to the strike of the range in the study area. Major faults (STF, MCT, MBT and MFT) are shown. Modern (black circles) and glacial-era (white squares) equilibrium-line altitudes define northward-rising gradients with a steeper gradient during glacial times. **b**, Monsoon precipitation projected against maximum topography through the

study area. Monsoon rainfall is ~ 10 -fold greater on the southern Himalayan flank than north of the Annapurna crest. The spatial extent of **c** is indicated. **c**, Apatite fission-track ages plotted against mean topography. Between the MCT and the STF system, the apatite fission-track ages are consistently very young ($<1 \text{ Myr}$). No apparent gradient in ages exists across the Greater Himalaya. Mean hillslope angles systematically increase northward across the zone of uniform cooling ages.

across the Greater Himalaya are rapid and spatially invariant.

These apatite fission-track dates are as young as, or younger than, apatite fission-track dates reported for Nanga Parbat¹⁴ and the Southern Alps of New Zealand¹⁵, where denudation is interpreted to exceed 4 mm yr^{-1} . Although the Marsyandi samples come from valley bottoms where isotherms should be most compressed, observed cooling rates across the Greater Himalaya of $\sim 300^\circ\text{C Myr}^{-1}$ are, nonetheless, consistent with long-term erosion rates of $\geq 2\text{--}5 \text{ mm yr}^{-1}$.

The abrupt increase in cooling rates at the northern edge of the Lesser Himalaya coincides spatially with a doubling of monsoon precipitation (Fig. 2b) due to the orographic effects of Greater Himalayan topography. Such correlations appear consistent with the hypothesis of a coupling between precipitation, channel gradients, and erosion rates. In contrast, the absence of a trend in apatite fission-track ages across the Greater Himalaya, despite at least a fivefold change in precipitation across this same area (Fig. 2b and c), indicates that no direct precipitation–erosion linkage exists across this area.

Alternative interpretations might explain spatially invariant cooling/erosion rates. In contrast to the present, average Pleistocene rainfall could have had a weaker gradient across the range. During the past 1 Myr, climatic variations have been primarily due to glacial–interglacial fluctuations. In the absence of direct measures of past precipitation, the equilibrium-line altitude of glaciers can serve as a proxy for precipitation changes¹⁶. Whereas the regional gradient of modern equilibrium-line altitudes in the Marsyandi catchment exhibits a northward-rising slope of $\sim 17 \text{ m km}^{-1}$, the gradient more than doubled to $>36 \text{ m km}^{-1}$ during glacial times (Fig. 2a), suggesting amplification of the regional rainfall gradient during glacial times. Thus, relative to the southern Himalayan flank, the northern half of the Marsyandi catchment should have been drier during glaciations than at present. Such conditions appear even more contradictory to a direct precipitation–erosion rate linkage.

A high geothermal gradient due to fluxes of hydrothermal water towards valley bottoms could create apatite fission-track dates that are younger than expected for any given erosion rate¹⁷. Indeed, numerous hot springs occur along the Marsyandi transect. Apatite

fission-track dates from ridge crests (Table 1), however, indicate rapid cooling rates ($>100^\circ\text{C Myr}^{-1}$) at sites up to 2 km above the valley bottom, thereby supporting the conclusion of rapid erosion, irrespective of the role of hydrothermal fluids. Moreover, the spatial uniformity of young apatite fission-track dates along the Marsyandi in the Greater Himalaya, rather than the absolute erosion rate, is central to our interpretation. Neither alternative interpretation offers a viable explanation for the observed uniformity in cooling rates.

Mechanisms by which spatially uniform erosion is maintained across the Greater Himalaya, despite greatly decreased precipitation, can be inferred from the digital topography and glacial history of the region. Although the extent of Late Pleistocene glaciation in the Marsyandi is debated^{18,19}, glacial sculpting of the landscape is unmistakable. Glaciers have the potential to erode at rates $\geq 5 \text{ mm yr}^{-1}$ (ref. 20); such rates would be sufficient to generate erosion rates in the glaciated upper Marsyandi equivalent to fluvial incision rates on the southern Himalayan flank, despite lower precipitation in the north.

In non-glaciated parts of the landscape, systematic topographic changes suggest how spatially uniform erosion rates are sustained for both hill slopes and river channels. First, as precipitation decreases northward from the MCT (Fig. 2b), mean hillslope angles increase (Fig. 2c), despite roughly uniform bedrock strength within the Greater Himalayan crystalline rocks²¹. Such steepening is consistent with hill slopes staying near the threshold for failure by bedrock landslides²², where higher pore pressures promote landsliding: steeper slopes compensate lower pore pressures in drier areas. Second, river channels could also become steeper to offset decreasing rainfall. Within the Greater Himalaya, however, channel gradients are already steep (~ 0.35) and do not increase northward (Fig. 3). Nonetheless, assuming that width scales with discharge, Q , as $Q^{0.4}$, specific stream power decreases only $\sim 45\%$ across the Greater Himalaya, despite an 80% drop in rainfall. Thus, to the extent that erosion rates scale with specific stream power²³, narrower channels offset about half of the decrease in precipitation. Third, if hillslope erosion is spatially uniform, then more northerly rivers should have higher sediment concentrations. Because most Himalayan rivers have excess sediment-transport capacity, higher sediment concentrations should enhance the frequency of clasts hitting the bed, thereby augmenting erosion rates²⁴ in drier areas. Thus, steeper hill slopes, narrower rivers, higher sediment concentrations, and efficiently eroding glaciers could yield spatially uniform erosion, despite decreasing precipitation and constant channel slopes.

These new Himalayan data do not, however, explain why hill slopes steepen or uniform erosion exists within the Greater Himalaya, but not elsewhere (Fig. 2). Precipitation clearly is not the dominant control. Instead, we suggest that convergence between India and southern Tibet ($\sim 18 \text{ mm yr}^{-1}$; ref. 25) has caused persistent lateral and vertical transport of rock into the orogen. When sustained for several million years, the resultant orogen should evolve toward flux and topographic steady states, in which erosional effluxes balance the tectonic influx²⁶, while crustal thickness and mean topography remain temporally constant. A north-dipping, crustal-scale ramp is proposed to lie beneath the Greater Himalaya^{27,28}, where its dip angle and associated slip rate should dictate upward particle velocities, with respect to the geoid, in the overthrusting plate. A planar ramp^{7,27,28} could predict ramp-parallel transport of rocks at a uniform rate. When balanced by erosion, this prediction is consistent with the observed zone of uniform fission-track ages. We conclude, therefore, that Indo-Tibetan convergence combines with the geometry of overthrusting to strongly influence patterns of erosion across the Greater Himalaya. The shape of the landscape itself results from an interplay of surface processes with precipitation gradients such that erosion is sufficient to balance tectonic deformation. □

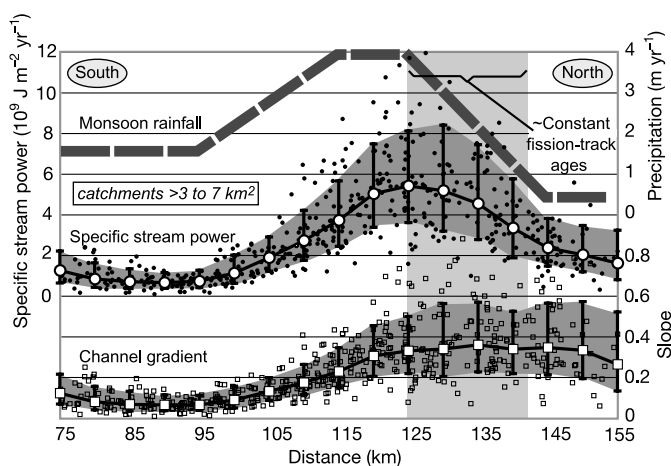


Figure 3 Plot of channel gradients, specific stream power, and precipitation gradient versus distance for Marsyandi tributary catchments ranging from 3 to 7 km^2 . The vertical grey swath indicates the zone in which fission-track ages average $\sim 0.5 \text{ Myr}$ and uniform, rapid erosion rates are inferred. Note that, although hillslope angles continue to steepen across this zone (Fig. 2c), channel slopes may have attained a threshold gradient (~ 0.35) in this same region. Whereas rainfall diminishes $\sim 80\%$ across this region, specific stream power only decreases 45% because narrower channels in drier areas focus the available stream power on a smaller channel perimeter.

Methods

Meteorology

A network of 14 meteorological stations was installed across the Annapurna range before the 1999 monsoon season, and expanded to 19 stations encompassing 28 rain gauges in 2000. Rainfall is totalled every 30 min. 'Look-down' distance rangiers and γ -ray loggers measure snow depth and total water content, respectively, once a day at high elevations (>2,500 m in the Greater Himalaya). Only liquid precipitation is measured in the Tibetan zone, such that the annual (but not the monsoon) total is underestimated here. The data presented here (Fig. 2b) represent monsoon averages based on the longest record available from each station.

Apatite fission-track dating

Following mineral separation, apatites were polished, etched and irradiated. Standard and induced track densities were determined on Brazil ruby muscovite external detectors (geometry factor 0.5), and fossil track densities were determined on internal mineral surfaces. Ages were calculated using $\bar{\epsilon} = 359 \pm 20$ for dosimeter glass CN-5. All ages are central ages and are reported with 1 σ errors. Long-term erosion rates are conservatively estimated on the basis of the fission-track age, and assuming a geothermal gradient of 100 °C km⁻¹ and an annealing temperature of 140 °C.

Topographic analysis

A 3-arcsec (~90 m) digital elevation model (DEM) is the basis of all topographic analyses. Hillslope angles are calculated at every pixel in the DEM based on a 3 × 3 pixel (~180 × 180 m) grid. Mean hillslope angles were extracted from a moving, 5-km-radius window centred on the Marsyandi River. Maximum, minimum and mean elevation (Fig. 2) were calculated along a 50-km-wide swath oriented perpendicular to the strike of the range and centred on the Marsyandi River (or the Nar-Phu River above its confluence with the Marsyandi).

Equilibrium-line altitude

Glacial areas were calculated from present and reconstructed ice margins mapped on aerial photographs, and transferred first to 1:50,000 scale topographic maps and then to the digital topography. Based on glacial hypsometry, equilibrium-line altitudes were estimated with an assumed accumulation-area ratio of 0.65. To avoid uncertainty introduced by avalanches on to glaciers from adjacent high peaks, 29 small glaciers (95% are <2.5 km²), lacking high headwalls, were analysed. The regional equilibrium-line altitude gradient shows little sensitivity to accumulation-area ratios ranging from 0.4 to 0.8.

Specific stream power

Analysis was focused on catchments ranging from 3 to 7 km² within the non-glaciated part (<4,200 m elevation) of the study area. These basins drain approximately half of the landscape and are sufficiently large to be fluvial, as opposed to colluvial/debris flow, channels. Monsoon rainfall was smoothed across the meteorological network to define an average precipitation gradient perpendicular to the strike of the topography. This gradient was then extrapolated parallel to strike across the study area. For each river segment ≥500 m long, channel gradients (S) were extracted from the DEM, and discharge (Q) was calculated as the product of upstream area and rainfall. Discharge is overestimated because all rainfall is assumed to enter channels. Channel width (W) is calculated as $10^{-2} Q^{0.4}$. Specific stream power (in GJ m⁻² yr⁻¹) is calculated as $\rho_w g Q S / W$, where ρ_w is the density of water and g is gravitational acceleration. Channel gradients and specific stream power are binned every 5 km.

Received 20 June; accepted 4 November 2003; doi:10.1038/nature02187.

- Koons, P. O. The topographic evolution of collisional mountain belts: A numerical look at the Southern Alps, New Zealand. *Am. J. Sci.* **289**, 1041–1069 (1989).
- Beaumont, C., Jamieson, R. A., Nguyen, M. H. & Lee, B. Himalayan tectonics explained by extrusion of a low-viscosity crustal channel coupled to focused surface denudation. *Nature* **414**, 738–742 (2001).
- Willett, S. D. Orogeny and orography: The effects of erosion on the structure of mountain belts. *J. Geophys. Res.* **104**, 28957–28982 (1999).
- Hodges, K. V. Tectonics of the Himalaya and southern Tibet from two perspectives. *Geol. Soc. Am. Bull.* **112**, 324–350 (2000).
- Tapponnier, P. et al. Oblique stepwise rise and growth of the Tibetan Plateau. *Science* **394**, 1671–1677 (2001).
- Burchfiel, B. D. et al. The South Tibetan detachment system, Himalayan orogen: Extension contemporaneous with and parallel to shortening in a collisional mountain belt. *Geol. Soc. Am. Spec. Pap.* **269**, 1–41 (1992).
- Harrison, T. M. et al. A late Miocene-Pliocene origin for the central Himalayan inverted metamorphism. *Earth Planet. Sci. Lett.* **146**, E1–E7 (1997).
- Catlos, E. J. et al. Geochronologic and thermobarometric constraints on the evolution of the Main Central Thrust, central Nepal Himalaya. *J. Geophys. Res.* **106**, 16177–16204 (2001).
- Robinson, D. M. et al. Kinematic model for the Main Central thrust in Nepal. *Geology* **31**, 359–362 (2003).
- Hurtado, J. M. Jr, Hodges, K. V. & Whipple, K. Neotectonics of the Thakkhola Graben and implications for Recent activity on the South Tibetan fault system in the central Nepal Himalaya. *Geol. Soc. Am. Bull.* **113**, 222–240 (2001).
- Dodson, M. H. in *Lectures in Isotope Geology* (eds Jaeger, E. & Hunziker, C. J.) 194–202 (Springer, New York, 1979).
- Stüwe, K., White, L. & Brown, R. The influence of eroding topography on steady-state isotherms. Application to fission track analysis. *Earth Planet. Sci. Lett.* **124**, 63–74 (1994).
- Willett, S. D. & Brandon, M. T. On steady states in mountain belts. *Geology* **30**, 175–178 (2002).
- Zeitler, P. K. Cooling history of the NW Himalaya, Pakistan. *Tectonics* **4**, 127–151 (1985).

- Tippett, J. M. & Kamp, P. J. J. Fission track analysis of the Late Cenozoic vertical kinematics of continental Pacific crust, South Island, New Zealand. *J. Geophys. Res.* **98**, 16119–16148 (1993).
- Porter, S. C. Some geological implication of average Quaternary glacial conditions. *Quat. Res.* **32**, 245–261 (1989).
- Ehlers, T. A., Armstrong, P. A. & Chapman, D. S. Normal fault thermal regimes and the interpretation of low-temperature thermochronometers. *Phys. Earth Planet. Inter.* **126**, 179–194 (2001).
- Fort, M. B. Glacial extension and catastrophic dynamics along the Annapurna Front, Nepal Himalaya. *Göttinger Geogr. Abh.* **81**, 105–121 (1986).
- Duncan, C. C., Klein, A. J., Masek, J. G. & Isacks, B. L. Comparison of Late Pleistocene and modern glacier extents in central Nepal based on digital elevation data and satellite imagery. *Quat. Res.* **49**, 241–254 (1998).
- Hallet, B., Hunter, L. & Bogen, J. Rates of erosion and sediment evacuation by glaciers: A review of field data and their implications. *Glob. Planet. Change* **12**, 213–235 (1996).
- Hodges, K. V., Parrish, R. R. & Searle, M. P. Tectonic evolution of the central Annapurna Range, Nepalese Himalayas. *Tectonics* **15**, 1264–1291 (1996).
- Burbank, D. W. et al. Bedrock incision, rock uplift, and threshold hillslopes in the northwestern Himalaya. *Nature* **379**, 505–510 (1996).
- Whipple, K. E., Kirby, E. & Brocklehurst, S. H. Geomorphic limits to climate-induced increases in topographic relief. *Nature* **401**, 39–43 (1999).
- Sklar, L. & Dietrich, W. E. Sediment supply, grain size and rock strength controls on rates of river incision into bedrock. *Geology* **29**, 1087–1090 (2001).
- Wang, Q. et al. Present-day crustal deformation in China constrained by Global Positioning System measurements. *Science* **294**, 574–577 (2001).
- Willett, S. D., Slingerland, R. & Hovius, N. Uplift, shortening, and steady state topography in active mountain belts. *Am. J. Sci.* **301**, 455–485 (2001).
- Seeber, L. & Gornitz, V. River profiles along the Himalayan arc as indicators of active tectonics. *Tectonophysics* **92**, 335–367 (1983).
- Pandey, M. R., Tandukar, R. P., Avouac, J. P., Lave, J. & Massot, J. P. Interseismic strain accumulation on the Himalayan crustal ramp (Nepal). *Geophys. Res. Lett.* **22**, 751–754 (1995).

Acknowledgements This work benefited from discussions with K. Hodges, J. Lavé, A. Heimsath, K. Whipple, P. Koons, M. Brandon and T. Ehlers. We thank P. Molnar and C. Beaumont for comments and suggestions. Logistical support from Himalayan Experience and the Nepalese Department of Hydrology and Meteorology is gratefully acknowledged. This work was funded by the NSF Continental Dynamics program and by NASA.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.W.B. (burbank@crustal.ucsb.edu).

Undesirable evolutionary consequences of trophy hunting

David W. Coltman¹, Paul O'Donoghue¹, Jon T. Jorgenson², John T. Hogg³, Curtis Strobeck⁴ & Marco Festa-Bianchet⁵

¹Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK

²Alberta Department of Sustainable Development, Fish and Wildlife Division, Box 1059, Canmore, Alberta T0L 0M0, Canada

³Montana Conservation Science Institute, Missoula, Montana 59803, USA

⁴Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9, Canada

⁵Département de biologie, Université de Sherbrooke, Sherbrooke, Québec J1K 2R1, Canada

Phenotype-based selective harvests, including trophy hunting, can have important implications for sustainable wildlife management if they target heritable traits^{1–3}. Here we show that in an evolutionary response to sport hunting of bighorn trophy rams (*Ovis canadensis*) body weight and horn size have declined significantly over time. We used quantitative genetic analyses, based on a partly genetically reconstructed pedigree from a 30-year study of a wild population in which trophy hunting targeted rams with rapidly growing horns⁴, to explore the evolutionary response to hunter selection on ram weight and horn size. Both traits were highly heritable, and trophy-harvested rams were of significantly higher genetic 'breeding value' for weight and horn size than rams that were not harvested. Rams of

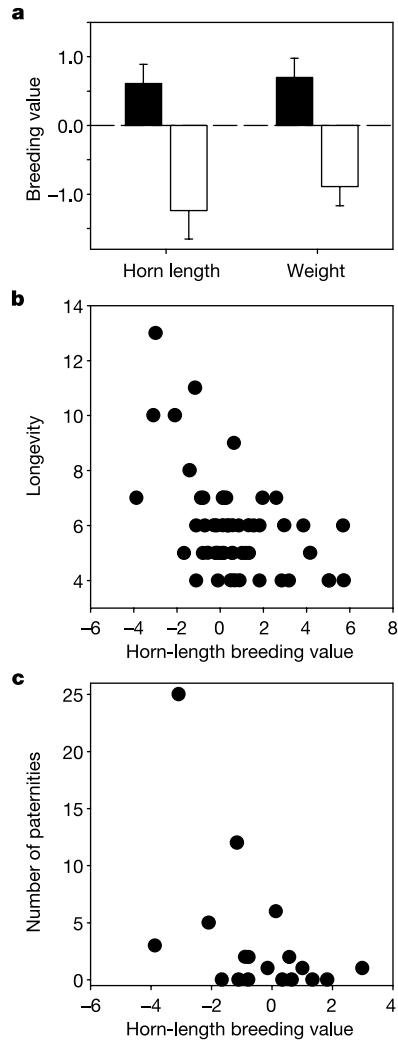


Figure 1 Selection against high-breeding-value rams imposed by trophy hunting. **a**, Breeding values (means $\pm$ s.e.m.) for horn length and weight of trophy-harvested rams (filled bars) and non-trophy-harvested rams (open bars). **b**, Relationship between the age at harvest for trophy-harvested rams and their breeding value. **c**, Relationship between the number of paternities assigned to trophy-harvested rams in their lifetime and their breeding value.

high breeding value were also shot at an early age, and thus did not achieve high reproductive success⁵. Declines in mean breeding values for weight and horn size therefore occurred in response to unrestricted trophy hunting, resulting in the production of smaller-horned, lighter rams, and fewer trophies.

Sport harvesting is one of the most pervasive and potentially intrusive human activities that affect game mammal populations globally⁶. Hunters are willing to pay large sums to hunt trophy mountain ungulates in various parts of the world, and many mountain sheep (*Ovis canadensis* and *O. dalli*) populations in North America are managed primarily to produce large-horned trophy rams for sport hunters. A world-class trophy ram is an extremely valuable commodity, and hunting permits have been auctioned for hundreds of thousands of dollars⁷. One sport hunter paid over Can\$1 million in 1998 and 1999 for special permits to hunt trophy rams in Alberta, Canada⁷. In many parts of North America, sport harvest of mountain sheep is often restricted only by the availability of rams whose horns reach a minimum size prescribed by regulations. Although the use of income generated from sport hunting towards enhancing and conserving mountain

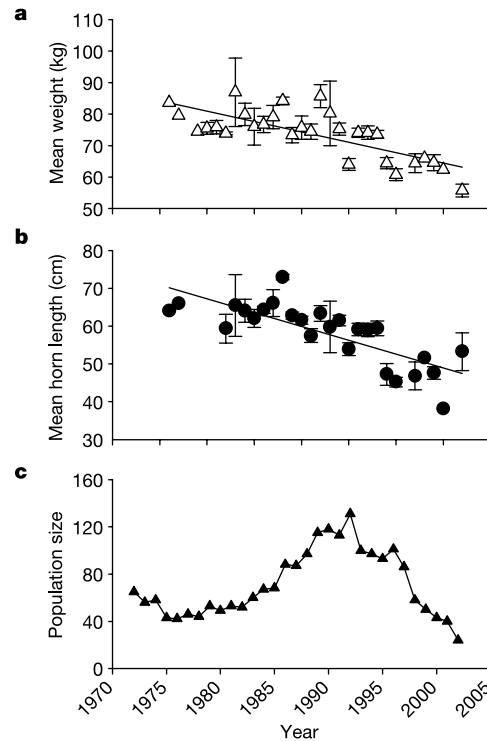


Figure 2 Observed changes in mean weight and horn length and in the population size from 1972 to 2002. **a**, Relationship between weight (mean $\pm$ s.e.m.) of 4-year-old rams and year ($N = 133$ rams). **b**, Relationship between horn length (mean $\pm$ s.e.m.) of 4-year-old rams and year ($N = 119$ rams). **c**, Changes in population size (taken as the number of ewes aged at least 2 years plus yearlings¹⁷) over time.

ungulate habitat can be seen in a positive light⁷, so far little attention has been paid to the potential evolutionary consequences, and hence the sustainability, of harvest regimes^{2,3}.

Wildlife management has traditionally focused on demographic and ecological factors that affect numbers and growth rates in harvested populations^{8–11}. However, the life-history changes experienced by species subject to commercial fisheries strongly suggest that intensive harvesting practices can elicit an evolutionary response in wild stocks^{12–15}. Experimental size-selective harvesting treatments on an exploited fish demonstrated evolutionary effects on somatic growth and population productivity in the opposite direction of the size bias of the harvest¹³. Recent reviews have called attention to the potential selective effects of sport hunting on wild ungulates, in which large-horned or large-antlered males are selectively targeted^{2,3}. The increased frequency of tuskless elephants in many African populations has also been suggested to have occurred in response to selective ivory poaching¹⁶. Here we use data from the long-term study of a harvested bighorn sheep population at Ram Mountain, Alberta, Canada, to investigate the evolutionary consequences of more than 30 years of selective hunting of trophy rams.

Fifty-seven rams have been shot at Ram Mountain since 1975, or about 40% of the rams legally available for harvest in each year (see Methods), for a yearly harvest of between zero and six rams¹⁷. Most trophy-harvested rams were shot before reaching 8 years of age (45 of 57 rams), and nine were shot as early as the age of 4 years. In bighorn sheep, much of the total horn length is added from the ages of 2 to 4 years, and at Ram Mountain the probability of a ram being shot before the age of 6 years is positively correlated with cumulative horn growth over this interval⁴. 'Animal model'¹⁸ quantitative genetic analysis of 395 horn-length and 447 weight measurements taken from 192 rams at ages 2, 3 and 4 years from 1971 to 2002 revealed narrow-sense heritabilities of 0.69 ± 0.10 and 0.41 ± 0.11

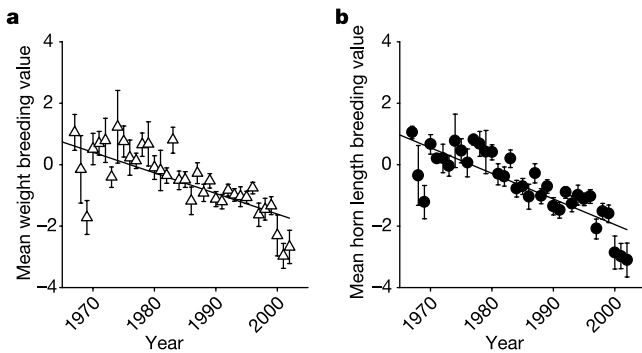


Figure 3 Changes in the mean breeding value of cohorts born between 1967 and 2002. **a**, Relationship between breeding value (mean $\pm$ s.e.m.) for weight and year of birth ($N = 783$ individuals). **b**, Relationship between breeding value (mean $\pm$ s.e.m.) for horn length and year of birth ($N = 783$ individuals).

(means $\pm$ s.e.m.), respectively (see Methods), and a strong positive additive genetic correlation between the two ($+0.84 \pm 0.10$). Comparison of expected genetic 'breeding values' (twice the expected deviation of an individual's offspring phenotype from the population mean owing to the additive effect of the offspring's inherited genes¹⁸) extracted from this model (Fig. 1a) indicates that hunters selectively harvest rams with high breeding values for horn length (trophy-harvested mean, $+0.61 \pm 0.28$; non-harvested mean, -1.24 ± 0.48 ; t -test: $t_{148} = -4.16$, $P < 0.001$) and weight (trophy-harvested mean, $+0.70 \pm 0.28$; non-harvested mean, -0.89 ± 0.48 ; t -test: $t_{148} = -3.26$, $P = 0.0014$).

Within seasons, mating success in bighorn sheep increases with dominance rank¹⁹, age and horn length⁵. The positive effect of large horns on mating success increases from about 6 years of age⁵, when rams are capable of defending oestrous ewes during the rut. The age at which a high-breeding-value ram is harvested is therefore likely to have an important impact on the number of offspring he can sire. We found a negative relationship between the age at which a trophy-harvested ram was shot and his breeding value for horn length (generalized linear model (GLM) with Poisson errors: $\chi^2_{(1)} = 4.64$, $P = 0.031$; Fig. 1b) but not for weight (GLM: $\chi^2_{(1)} = 1.80$, $P = 0.18$; data not shown). Trophy-harvested rams with high breeding values for body and horn size were therefore less likely to reach the ages at which they achieve high rates of paternity in this population⁵. As a consequence, there was a negative relationship between breeding value for horn length and lifetime mating success, measured as the number of paternities assigned over their lifetime, among trophy-harvested rams (GLM with negative binomial error: $\chi^2_{(1)} = 8.56$, $P = 0.0034$; Fig. 1c). The mean sire breeding value of individuals fathered by trophy-harvested rams was therefore significantly less than zero for both weight (one-sample t -test: mean = -2.41 , s.e.m. = 0.37 , $t_{59} = -6.50$, $P < 0.001$) and horn length (mean = -1.84 , s.e.m. = 0.19 , $t_{59} = -9.68$, $P < 0.001$). The mean sire breeding value of individuals fathered by rams that died a natural death was also significantly less than zero for both weight (one-sample t -test: mean = -1.24 , s.e.m. = 0.17 , $t_{182} = -7.14$, $P < 0.001$) and horn length (mean = -2.10 , s.e.m. = 0.16 , $t_{182} = -20.43$, $P < 0.001$). The low breeding values of rams not harvested (Fig. 1a) and the reduced longevity and potential reproductive output of the higher-quality trophy-harvested rams (Fig. 1b, c) combine to suggest that the selection imposed by trophy hunting had a negative impact on the evolutionary trajectory of horn length and body weight in this population during our study.

Is there evidence of a response to selective harvesting at the population level? Significant declines in both ram weight (linear mixed-effect model including year of birth and individual as a

random effects, and age, time and resource index as fixed effects: $\beta_{\text{time}} = -0.30$, s.e.m. = 0.09 , $t_{25} = -3.42$, $P = 0.0021$) and horn length (linear mixed-effect model including year of birth and individual as a random effects, and age, time and resource index as fixed effects: $\beta_{\text{time}} = -0.35$, s.e.m. = 0.12 , $t_{23} = -2.97$, $P = 0.0068$) were observed over the course of the study (Fig. 2a, b) after controlling for environmental effects such as population density (Fig. 2c) using an index of resource availability (see Methods; weight: $\beta_{\text{resources}} = 0.81$, s.e.m. = 0.17 , $t_{25} = 4.72$, $P < 0.001$; horn length: $\beta_{\text{resources}} = 0.72$, s.e.m. = 0.22 , $t_{23} = 3.32$, $P = 0.0030$). These are very rapid rates of phenotypic change²⁰, corresponding to $-0.30/12.9 = -0.023$ and $-0.35/13.6 = -0.026$ standard deviations per year, or -0.14 and -0.15 haldanes (ref. 20) assuming a generation time of 6 years. Analyses of breeding values are consistent with genetically based responses (Fig. 3). Declines in breeding value (see Methods) were observed for both ram weight (linear mixed-effect model including year of birth as a random effect, and time and resource index as fixed effects: $\beta_{\text{resources}} = 0.037$, s.e.m. = 0.025 , $t_{33} = 1.49$, $P = 0.15$; $\beta_{\text{time}} = -0.071$, s.e.m. = 0.012 , $t_{33} = -6.02$, $P < 0.001$) and horn length (linear mixed-effect model including year of birth as a random effect, and time and resource index as fixed effects: $\beta_{\text{resources}} = 0.050$, s.e.m. = 0.024 , $t_{33} = 2.08$, $P = 0.045$; $\beta_{\text{time}} = -0.075$, s.e.m. = 0.011 , $t_{33} = -6.76$, $P < 0.001$). Such declines in breeding value over time are indicative of a microevolutionary response to selection²¹ in the Ram Mountain population.

Unrestricted harvesting of trophy rams has thus contributed to a decline in the very traits that determine trophy quality. Hunters have selectively targeted rams of high genetic quality before their reproductive peak, depleting the genes that confer rapid early body and horn growth. Wildlife harvesting that is selective and sufficiently severe might elicit an undesired evolutionary response when the target trait is heritable. There might also be unexpected effects on genetically correlated traits, such as female body weight or disease resistance²², that could result in further genetic deterioration of harvested populations as anthropogenic selection pushes traits away from their naturally selected optima. Because such changes will be extremely difficult to reverse, wildlife managers must consider the genetic effects and the evolutionary implications of alternative harvest strategies^{2,3}. The move to adopt a 'full curl' restriction in parts of Alberta in 1996, which limits harvest to rams with horns whose tip extends beyond the tip of the nose, is one strategy to minimize further deterioration of the genetic quality of bighorn sheep. □

Methods

Population and study site

The bighorn sheep population on Ram Mountain, Alberta, Canada (52°N , 115°W ; elevation 1,080–2,170 m) has been monitored closely since 1971 (refs 17, 23). Immigration to Ram Mountain from the main species range has not been documented, and is probably rare because of isolation of the population by about 30 km of coniferous forest. Each year, sheep were captured in a corral trap baited with salt from late May to early October, and marked with coloured plastic ear tags or canvas collars for individual identification. Adult rams were captured once or twice in most summers from early June to mid-July. At each capture, sheep were weighed to the nearest 250 g with a Deteco spring scale. Horn length along the outside curvature was measured with tape. The longer of the left and right horn measurements was used, because rams can have a varying amount of horn removed by wear. For further details on field methods see refs 17, 23 and 24.

Bighorn males on Ram Mountain can be legally harvested by Alberta resident hunters from late August to the end of October. Until 1996, rams with horns describing at least four-fifths of a curl ('trophy' rams) could be harvested by any hunter holding a trophy sheep licence¹⁷. As any resident could purchase a licence, the harvest was limited only by the availability of trophy rams. A change in regulations in 1996 limited harvest to 'full-curl' rams. Consequently, only three rams have been shot since 1996. Individual weight and horn length measurements from rams captured between 1971 and 2002 were adjusted to 5 June (ref. 24). Because the youngest age at which rams were shot by hunters was 4 years, we used weight and horn length data from ages 2, 3 and 4 years to avoid bias due to hunter selection.

Pedigree reconstruction

Maternity was known from field observations for 709 of the 894 (79.3%) marked sheep

whose fates have been followed since 1971. Tissue sampling for DNA analyses started in 1988. Blood samples were taken from all captured sheep until 1993 and stored in preservative at -20°C . Sampling resumed in 1997, when hair samples were taken from all captured sheep by plucking 50–100 hairs including roots from the back or flank. Hairs were kept either in paper envelopes or plastic bags containing about 5 g of silica at room temperature. From 1998 to 2002, a tissue sample from each captured sheep was taken from the ear with an 8-mm punch. Ear tissue was kept at -20°C in a solution of 20% dimethylsulphoxide saturated with NaCl. We sampled 433 marked individuals over the course of the study.

DNA was extracted from blood with a standard phenol–chloroform method, and from either 20–30 hairs including follicles or about 5 mg of ear tissue, using the QIAamp tissue extraction kit (Qiagen Inc., Mississauga, Ontario). Polymerase chain reaction amplification at 20 ungulate-derived microsatellite loci, 15 as described previously⁵ plus MCM527, BM4025, MAF64, OarFCB193 and MAF92 (refs 25, 26), and fragment analysis were performed as described elsewhere⁵. After correction for multiple comparisons, we found no evidence for allelic or genotypic disequilibria at or among these 20 loci.

Paternity of 241 individuals was assigned by using the likelihood-based approach described in CERVUS²⁷ at a confidence level of more than 95% with input parameters given in ref. 5. After paternity analysis, we used KINSHIP²⁸ to identify 31 clusters of 104 paternal half-sibs among the unassigned offspring. A paternal half-sibship consisted of all pairs of individuals of unassigned paternity that were identified in the KINSHIP analysis as having a likelihood ratio of the probability of a paternal half-sib relationship versus unrelated with an associated $P < 0.05$ (ref. 28). Members of reconstructed paternal half-sibships were assigned a common unknown paternal identity for the animal model analyses. Paternal identity links in the pedigree were therefore defined for 345 individuals.

Animal model analyses

Breeding values, genetic variance components and heritabilities were estimated by using a multiple trait restricted-estimate maximum-likelihood (REML) model implemented by the programs PEST²⁹ and VCE³⁰. An animal model was fitted in which the phenotype of each animal was broken down into components of additive genetic value and other random and fixed effects: $y = Xb + Za + Pc + e$, where y was a vector of phenotypic values, b was a vector of fixed effects, a and c were vectors of additive genetic and permanent environmental, e was a vector of residual values, and X , Z and P were the corresponding design matrices relating records to the appropriate fixed or random effects¹⁸. Fixed effects included age (factor) and the average weight of yearling ewes in the year of measurement (covariate), which is a better index of resource availability than population size because it accounts for time-lagged effects⁴. The permanent environmental effect grouped repeated observations on the same individual to quantify any remaining between-individual variance over and above that due to additive genetic effects, which would be due to maternal or other long-term environmental and non-additive genetic effects.

The total phenotypic variance (V_p) was therefore partitioned into three components: the additive genetic variance (V_a), the permanent environmental variance (V_e) and the residual variance (V_r), thus: $V_p = V_a + V_e + V_r$. Heritability was calculated as $h^2 = V_a/V_p$. The VCE³⁰ program returns standard errors on all variance components and ratios. Best linear unbiased predictors of individual breeding values were quantified by using REML estimates of the variance components obtained with PEST²⁹. All statistical tests were conducted in SPLUS 6.1.

Received 11 August; accepted 17 October 2003; doi:10.1038/nature02177.

- Thelen, T. H. Effects of harvest on antlers of simulated populations of elk. *J. Wild. Mgmt* **55**, 243–249 (1991).
- Harris, R. B., Wall, W. A. & Allendorf, F. W. Genetic consequences of hunting: what do we know and what should we do? *Wildl. Soc. Bull.* **30**, 634–643 (2002).
- Festa-Bianchet, M. in *Animal Behavior and Wildlife Conservation* (eds Apollonio, M. & Festa-Bianchet, M.) 191–207 (Island Press, Washington DC, 2003).
- Festa-Bianchet, M., Coltman, D. W., Turelli, L. & Jorgenson, J. T. Relative allocation to horn and body growth in bighorn rams varies with resource availability. *Behav. Ecol.* (in the press).
- Coltman, D. W., Festa-Bianchet, M., Jorgenson, J. T. & Strobeck, C. Age-dependent sexual selection in bighorn rams. *Proc. R. Soc. Lond. B* **269**, 165–172 (2002).
- Milner-Gulland, E. J. *et al.* Conservation—reproductive collapse in saiga antelope harems. *Nature* **422**, 135 (2003).
- Marty, S. Sacrificial ram. *Can. Geographic* November/December, 37–50 (2002).
- Clutton-Brock, T. H., Coulson, T. N., Milner-Gulland, E. J., Thomson, D. & Armstrong, H. M. Sex differences in emigration and mortality affect optimal management of deer populations. *Nature* **415**, 633–637 (2002).
- Ginsberg, J. R. & Milner-Gulland, E. J. Sex-biased harvesting and population dynamics in ungulates—implications for conservation and sustainable use. *Conserv. Biol.* **8**, 157–166 (1994).
- Langvatn, R. & Loison, A. Consequences of harvesting on age structure, sex ratio and population dynamics of red deer *Cervus elaphus* in central Norway. *Wildl. Biol.* **5**, 213–223 (1999).
- Laurian, C., Ouellet, J. P., Courtois, R., Breton, L. & St-Onge, S. Effects of intensive harvesting on moose reproduction. *J. Appl. Ecol.* **37**, 515–531 (2000).
- Ashley, M. V. *et al.* Evolutionarily enlightened management. *Biol. Conserv.* **111**, 115–123 (2003).
- Conover, D. O. & Munch, S. B. Sustaining fisheries yields over evolutionary time scales. *Science* **297**, 94–96 (2002).
- Palumbi, S. R. Evolution—humans as the world's greatest evolutionary force. *Science* **293**, 1786–1790 (2001).
- Law, R. Fishing, selection, and phenotypic evolution. *ICES J. Mar. Sci.* **57**, 659–668 (2000).
- Jachmann, H., Berry, P. S. M. & Imae, H. Tusklessness in African elephants—a future trend. *Afr. J. Ecol.* **33**, 230–235 (1995).
- Jorgenson, J. T., Festa-Bianchet, M. & Wishart, W. D. Effects of population density on horn development in bighorn rams. *J. Wildl. Mgmt* **62**, 1011–1020 (1998).

- Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Massachusetts, Sinauer, Sunderland, 1998).
- Hogg, J. T. & Forbes, S. H. Mating in bighorn sheep: frequent male reproduction via a high-risk 'unconventional' tactic. *Behav. Ecol. Sociobiol.* **41**, 33–48 (1997).
- Hendry, A. P. & Kinnison, M. T. Perspective: The pace of modern life: measuring rates of contemporary microevolution. *Evolution* **53**, 1637–1653 (1999).
- Merilä, J., Kruuk, L. E. B. & Sheldon, B. C. Cryptic evolution in a wild bird population. *Nature* **412**, 76–79 (2001).
- Coltman, D. W., Pilkington, J., Kruuk, L. E. B., Wilson, K. & Pemberton, J. M. Positive genetic correlation between parasite resistance and body size in a free-living ungulate population. *Evolution* **55**, 2116–2125 (2001).
- Jorgenson, J. T., Festa-Bianchet, M. & Wishart, W. D. Harvesting bighorn ewes—consequences for population-size and trophy ram production. *J. Wildl. Mgmt* **57**, 429–435 (1993).
- Festa-Bianchet, M., Jorgenson, J. T., King, W. J., Smith, K. G. & Wishart, W. D. The development of sexual dimorphism: seasonal and lifetime mass changes in bighorn sheep. *Can. J. Zool.* **74**, 330–342 (1996).
- Crawford, A. M. *et al.* An autosomal genetic linkage map of the sheep genome. *Genetics* **140**, 703–724 (1995).
- Slate, J. *et al.* Bovine microsatellite loci are highly conserved in red deer (*Cervus elaphus*), sika deer (*Cervus nippon*) and Soay sheep (*Ovis aries*). *Anim. Genet.* **29**, 307–315 (1998).
- Marshall, T. C., Slate, J., Kruuk, L. E. B. & Pemberton, J. M. Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol.* **7**, 639–655 (1998).
- Goodnight, K. F. & Queller, D. C. Computer software for performing likelihood tests of pedigree relationship using genetic markers. *Mol. Ecol.* **8**, 1231–1234 (1999).
- Groeneveld, E., Kovac, M., Wang, T. L. & Fernando, R. L. Computing algorithms in a general purpose BLUP package for multivariate prediction and estimation. *Arch. Anim. Breed.* **15**, 399–412 (1992).
- Groeneveld, E. *User's Guide: REML VCE, a Multivariate Multi-Model Restricted Maximum Likelihood (Co)variance Estimation Package, Version 3.2* (Institute of Animal Husbandry and Animal Behaviour, Federal Research Center of Agriculture (FAL) (Mariensee, Germany, 1995).

Acknowledgements We thank the many students, colleagues, volunteers and assistants that contributed to this research over the past 30 years. B. Wishart initiated the Ram Mountain project. Our research was funded by the Alberta Conservation Association, Alberta Fish and Wildlife Division, Alberta Recreation, Sports, Parks and Wildlife Foundation, Eppley Foundation for Research, Foundation for North American Wild Sheep, National Geographic Society, Natural Environment Research Council (UK), Natural Sciences and Engineering Research Council of Canada, Rocky Mountain Elk Foundation (Canada), and the Université de Sherbrooke. We are grateful for the logistical support of the Alberta Forest Service.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.W.C. (d.coltman@sheffield.ac.uk).

The role of evolution in the emergence of infectious diseases

Rustum Antia¹, Roland R. Regoes¹, Jacob C. Koella² & Carl T. Bergstrom³

¹Department of Biology, Emory University, Atlanta, Georgia 30322, USA

²Laboratoire de Parasitologie Evolutive, Université Pierre et Marie Curie, 75252 Paris, France

³Department of Biology, University of Washington, Seattle, Washington 98195, USA

It is unclear when, where and how novel pathogens such as human immunodeficiency virus (HIV), monkeypox and severe acute respiratory syndrome (SARS) will cross the barriers that separate their natural reservoirs from human populations and ignite the epidemic spread of novel infectious diseases. New pathogens are believed to emerge from animal reservoirs when ecological changes increase the pathogen's opportunities to enter the human population¹ and to generate subsequent human-to-human transmission². Effective human-to-human transmission requires that the pathogen's basic reproductive number, R_0 , should exceed one, where R_0 is the average number of secondary infections arising from one infected individual in a completely susceptible population³. However, an increase in R_0 , even when insufficient to generate an epidemic, nonetheless increases the number of subsequently infected individuals. Here we show that,

as a consequence of this, the probability of pathogen evolution to $R_0 > 1$ and subsequent disease emergence can increase markedly.

The emergence of a disease combines two elements: the introduction of the pathogen into the human population and its subsequent spread and maintenance within the population. Ecological factors such as human behaviour can influence both of these elements, and consequently ecology has been recognized to have an important role in the emergence of disease^{1,2,4}. In contrast, evolutionary factors including the adaptation of the pathogen to growth within humans and the subsequent transmission of the pathogen between humans are mostly considered in terms of changes in the virulence of the pathogen, and are often thought to have a lesser role in the initial emergence of pathogens⁴. One exception⁵ suggests that immunocompromised individuals might provide “stepping stones” for the evolution of pathogens.

The successful emergence of a pathogen requires R_0 to exceed one in the new host. Only then can an introduction trigger emergence². (Here we use R_0 to refer to spread in human populations, not in the natural reservoir.) If R_0 for a potential pathogen exceeds one, this scenario represents an epidemic waiting to happen. By contrast, when R_0 is initially less than one, infections will inevitably die out and there will be no epidemic unless genetic or ecological changes drive R_0 above one.

There are a number of ways in which R_0 can increase. Ecological changes such as changes in host density or behaviour can increase R_0 , as can genetic changes in the pathogen population or in the population of its new host. Genetic changes in the pathogen can arise either through ‘coincidental’ processes such as neutral drift or coevolution of the pathogen and its reservoir host, or through adaptive evolution of the pathogen during chains of transmission in humans. Genetic changes of the new host might be more likely for domesticated or endangered species than for humans.

Here we show that factors, such as ecological changes, that increase the R_0 value of the potential pathogen to a level not sufficient to cause an epidemic (that is, R_0 remains less than one) can greatly increase the length of the stochastic chains of disease transmission. These long transmission chains provide an opportunity for the pathogen to adapt to human hosts, and thus for the disease to emerge.

Our model is illustrated in Fig. 1. Introductions occur stochastically from the natural reservoir of the pathogen, and each primary case is followed by stochastic transmission that generates a variable number of subsequent infections in the human population. We

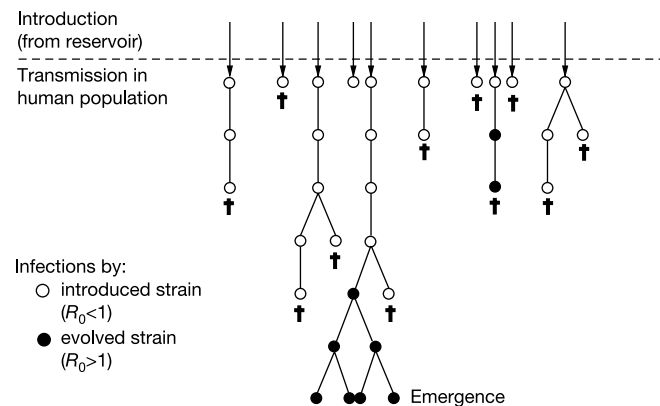


Figure 1 Schematic for the emergence of an infectious disease. Introductions from the reservoir are followed by chains of transmission in the human population. Infections with the introduced strain (open circles) have a basic reproductive number $R_0 < 1$. Pathogen evolution generates an evolved strain (filled circles) with $R_0 > 1$. The infections caused by the evolved strain can go on to cause an epidemic. Daggers indicate no further transmission.

assume that the number of secondary cases follows a Poisson distribution with a mean equal to R_0 . Each introduction thus forms branched chains of transmission, which stutter to extinction if $R_0 < 1$, and the pathogen cannot evolve ($\mu = 0$). The probability that the pathogen evolves and a secondary infection is caused by the mutant is equal to μ for each of the secondary infections (we note that μ incorporates not only the pathogen's mutation rate, but also its dynamics within the host and its transmissibility). We use multi-type branching processes^{6–9} to describe the initial spread of the infection, incorporating the evolution of the pathogen.

In the simplest case (see Fig. 2a) only one mutation is required for the R_0 of the evolved pathogen to exceed one. The probability that a single introduction evolves, causing one (or more) infections with the evolved pathogen (a filled circle in Fig. 1) before it goes extinct, depends very strongly on the R_0 of the introduced pathogen, particularly for low μ values as R_0 approaches 1. This probability is approximately linearly dependent on the rate of evolution μ .

The probability that the introduction leads to an epidemic (the ‘probability of emergence’) depends on the probability of evolution and the probability that the evolved infections do not go extinct due to stochastic effects. In Fig. 2b we plot the probability of emergence for three different R_0 values of the evolved strain. The probability of emergence approaches the probability of evolution when R_0 of the evolved strain is large, and is lower when the R_0 of the evolved strain

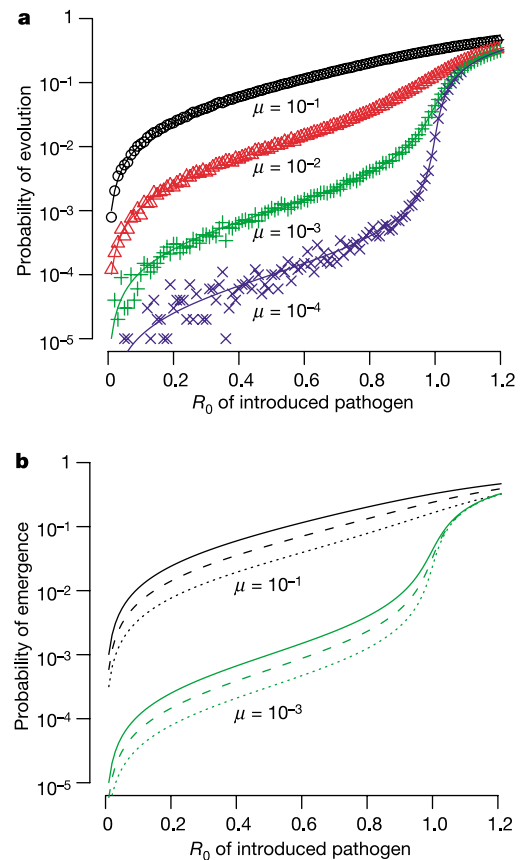


Figure 2 One-step evolution. A single change is required for the pathogen to evolve to $R_0 > 1$. **a**, The probability that an introduction leads to an infection with an evolved strain of the pathogen (filled circle in Fig. 1) is highly sensitive to R_0 , and is approximately linearly dependent on the mutation rate μ . Lines correspond to numerical solutions to the branching process model (see Supplementary Information) and symbols correspond to Monte-Carlo simulations following 10^5 introductions. **b**, The probability of emergence per introduction depends on the R_0 value of the introduced pathogen and of the evolved pathogen. The solid, dashed and dotted lines correspond to the evolved pathogen having an R_0 of 1.0, 1.5 and 1.2 respectively.

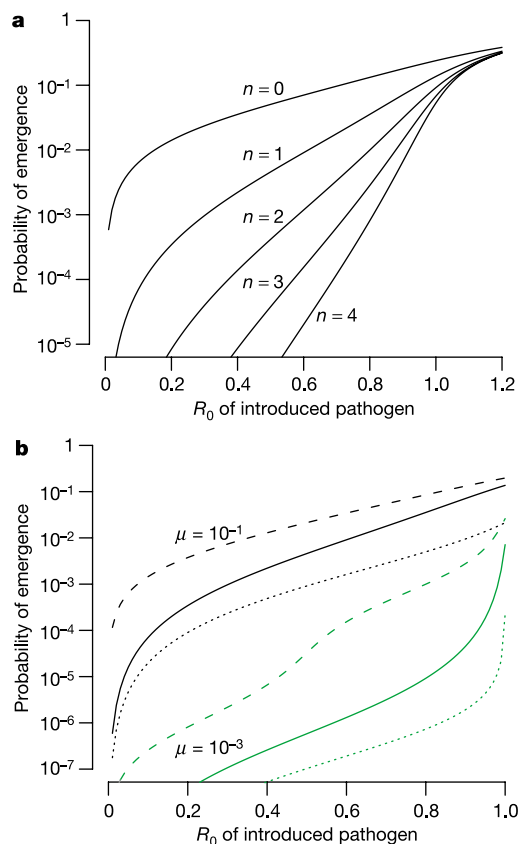


Figure 3 Multiple-step evolution. Here multiple evolutionary changes are required for evolution of the pathogen to have an $R_0 > 1$. **a**, Jackpot model with $\mu = 0.1$ and n intermediate changes each with R_0 equal to that of the introduced pathogen: increasing the number of steps (n) greatly decreases the probability of evolution, and makes it more sensitive to the R_0 of the introduced pathogen. **b**, Alternative multi-step models for the one-intermediate ($n = 1$) case. The jackpot model (solid line), additive model (dashed line) and fitness valley model (dotted line) are shown (see text for details).

is close to 1. We find that the probability of emergence depends most strongly on the R_0 of the introduced pathogen, increases approximately linearly with the mutation rate μ , and depends only modestly on the R_0 of the evolved pathogen.

We extend the simple one-step mutation model to consider the situation in which multiple evolutionary changes are required for the pathogen to attain $R_0 > 1$ in the human population. We begin with a simple scenario, which we call the jackpot model, where the R_0 of the pathogen with the intermediate mutations is the same as that of the introduced pathogen, and where only the addition of the final mutation results in an increase in R_0 to greater than one. As seen in Fig. 3, increasing the number of required evolutionary steps greatly reduces the probability of emergence and increases its sensitivity to changes in R_0 . The probability of emergence is approximately proportional to the mutation rate to the power of the number of evolutionary steps required (see Supplementary Information).

The fitness landscape on which evolution occurs is important in determining the outcome¹⁰. Figure 3b illustrates this for the case of a single intermediate type. As should be expected, changing the jackpot model to an additive model, where the fitness of the intermediate is the average of the fitness of the introduced strain and fully evolved strain, increases the probability that the pathogen evolves to $R_0 > 1$, whereas changing it to a fitness valley model, where the fitness of the intermediate is lower than the fitness of the introduced strain, decreases the probability of emergence.

The key characteristic distinguishing our model from the con-

ventional view is the R_0 of the introduced pathogen. In the conventional view the R_0 of these infections must be greater than one, whereas in the mechanism described here it is less than one, and evolution during the stochastic chains of transmission allows R_0 to increase above one. In the case of human infections such as HIV^{1,11,12}, SARS^{13–17} and (potentially) monkeypox^{18–21}, seroprevalence studies among groups at high risk of infection from the reservoir would allow evaluation of whether crossover events are usually dead ends, as we expect in our model, or whether they are associated with a large number of secondary cases. In the case of diseases emerging into non-human populations, such as the Nipah virus, which moved from bats to pigs^{22–24}, it may be possible to conduct additional tests involving controlled experimental infections to estimate the R_0 (in this case of the bat Nipah virus in pigs) and to determine whether it evolves during the course of chains of transmission. Such studies may additionally help to identify pathogens that have an ability to evolve rapidly and thus have a high potential for emergence.

The framework presented here has special relevance for pathogens that have been driven to extinction by vaccination. In the case of smallpox there are probably reservoirs of related zoonoses (such as, but not restricted to, monkeypox) from which smallpox may have originated. Although the R_0 of monkeypox in the human population is clearly less than one, there are occasional chains of transmission in the human population^{18–21}. As the level of herd immunity to smallpox wanes in the absence of continued vaccination, we expect an increase in R_0 of infections with monkeypox albeit to a level still less than one (the smallpox vaccine provides about 85% cross immunity against monkeypox). Our results suggest that this increase in the effective R_0 of monkeypox in the human population could markedly increase the probability of evolution of monkeypox, allowing it to emerge into a successful human pathogen (which, depending on the evolutionary trajectory followed, may be similar to or differ from smallpox).

The present study could be extended in a number of directions. These include explicitly incorporating the details of ecological interactions such as heterogeneity in the transmission in different areas and subpopulations^{2,25,26,27} and incorporating genetic diversity of the pathogen in its reservoir. Finally, we note that this framework can be applied to the more general problem of biological invasions^{28,29}. □

Methods

We describe the dynamics and evolution of emerging diseases as a multi-type branching process with the following probability-generating functions^{6–9}:

$$f_i(s_1, s_2, \dots, s_m) = \exp[-(1 - \mu)R_0^{(i)}(1 - s_i)] \exp[-\mu R_0^{(i)}(1 - s_{i+1})], \quad i = 1, \dots, m-1$$

$$f_m(s_1, s_2, \dots, s_m) = \exp[-R_0^{(m)}(1 - s_m)]$$

We calculated the extinction probabilities of the above process numerically. For details and definitions see Supplementary Information.

Received 6 July; accepted 25 September 2003; doi:10.1038/nature02104.

- Hahn, B. H., Shaw, G. M., De Cock, K. M. & Sharp, P. M. AIDS as a zoonosis: scientific and public health implications. *Science* **287**, 607–614 (2000).
- May, R. M., Gupta, S. & McLean, A. R. Infectious disease dynamics: What characterizes a successful invader? *Phil. Trans. R. Soc. Lond. B* **356**, 901–910 (2001).
- Anderson, R. & May, R. *Infectious Diseases of Humans: Dynamics and Control* 1st edn (Oxford Univ. Press, Oxford, 1991).
- Schrag, S. & Wiener, P. Emerging infectious disease: what are the relative roles of ecology and evolution? *Trends Ecol. Evol.* **10**, 319–324 (1995).
- Wallace, B. Can “stepping stones” form stairways? *Am. Nat.* **133**, 578–579 (1989).
- Athreya, K. B. & Ney, P. *Branching Processes* (Springer, Berlin/New York, 1972).
- Demetrius, L., Schuster, P. & Sigmund, K. Polynucleotide evolution and branching processes. *Bull. Math. Biol.* **47**, 239–262 (1985).
- Stivers, D. N., Kimmel, M. & Axelrod, D. E. A discrete-time, multi-type generational inheritance branching process model of cell proliferation. *Math. Biosci.* **137**, 25–50 (1996).
- Wilke, C. O. Probability of fixation of an advantageous mutant in a viral quasispecies. *Genetics* **163**, 467–474 (2003).
- Wright, S. The role of mutation, inbreeding, crossbreeding and selection in evolution. *Proc. 6th Int. Cong. Genet.* **1**, 356–366 (1932).
- Holmes, E. C. On the origin and evolution of the human immunodeficiency virus (HIV). *Biol. Rev.*

- Camb. Phil. Soc.* **76**, 239–254 (2001).
12. Peeters, M. *et al.* Risk to human health from a plethora of simian immunodeficiency viruses in primate bushmeat. *Emerg. Infect. Dis.* **8**, 451–457 (2002).
 13. Guan, Y. *et al.* Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* **302**, 276–278 (2003).
 14. Cyranoski, D. & Abbott, A. Virus detectives seek source of SARS in China's wild animals. *Nature* **423**, 467 (2003).
 15. Enserink, M. Infectious diseases. Clues to the animal origins of SARS. *Science* **300**, 1351 (2003).
 16. Riley, S. *et al.* Transmission dynamics of the etiologic agent of SARS in Hong Kong: impact of public health interventions. *Science* **300**, 1961–1966 (2003).
 17. Lipsitch, M. *et al.* Transmission dynamics and control of Severe Acute Respiratory Syndrome. *Science* **300**, 1966–1970 (2003).
 18. Heymann, D. L., Szczeniowski, M. & Esteves, K. Re-emergence of Monkeypox in Africa: a review of the past six years. *Br. Med. Bull.* **54**, 693–702 (1998).
 19. Fine, P. E., Jezek, Z., Grab, B. & Dixon, H. The transmission potential of monkeypox virus in human populations. *Int. J. Epidemiol.* **17**, 643–650 (1988).
 20. Breman, J. G. & Henderson, D. A. Poxvirus dilemmas—Monkeypox, smallpox, and biological terrorism. *N. Engl. J. Med.* **339**, 556–559 (1998).
 21. Hutin, Y. J. *et al.* Outbreak of human monkeypox, Democratic Republic of Congo, 1996 to 1997. *Emerg. Infect. Dis.* **7**, 434–438 (2001).
 22. Enserink, M. New virus fingered in Malaysian epidemic. *Science* **284**, 407, 409–410 (1999).
 23. Chua, K. B. *et al.* Nipah virus: a recently emergent deadly paramyxovirus. *Science* **288**, 1432–1435 (2000).
 24. Enserink, M. Emerging diseases. Malaysian researchers trace Nipah virus outbreak to bats. *Science* **289**, 518–519 (2000).
 25. May, R. & Anderson, R. Parasite–host coevolution. *Parasitology* **100**, S89–S101 (1990).
 26. Lloyd, A. L. & May, R. M. Spatial heterogeneity in epidemic models. *J. Theor. Biol.* **179**, 1–11 (1996).
 27. Dobson, A. & Foufopoulos, J. Emerging infectious pathogens of wildlife. *Phil. Trans. R. Soc. B.* **356**, 1001–1012 (2001).
 28. Elton, C. S. *Ecology of Invasions by Animals and Plants* (Methuen, London, 1958).
 29. Kornberg, H. & Williamson, M. *Quantitative Biology of the Ecology of Biological Invasions* (The Royal Society, London, 1987).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank S. Nichol, I. Longini and D. Krakauer for discussions and comments. We acknowledge support from the NIH and DFG.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.A. (rantia@emory.edu).

Plankton effect on cod recruitment in the North Sea

Grégory Beaugrand^{1,2}, Keith M. Brander³, J. Alistair Lindley², Sami Souissi¹ & Philip C. Reid²

¹CNRS, UMR 8013 ELICO, Station Marine, Université des Sciences et Technologies de Lille BP 80, 62930 Wimereux, France

²Sir Alister Hardy Foundation for Ocean Science, The Laboratory Citadel Hill, Plymouth PL1 2PB, UK

³ICES, Palaegade 2–4, 1261 Copenhagen K, Denmark

The Atlantic cod (*Gadus morhua* L.) has been overexploited in the North Sea since the late 1960s and great concern has been expressed about the decline in cod biomass and recruitment¹. Here we show that, in addition to the effects of overfishing¹, fluctuations in plankton have resulted in long-term changes in cod recruitment in the North Sea (bottom-up control). Survival of larval cod is shown to depend on three key biological parameters of their prey: the mean size of prey, seasonal timing and abundance. We suggest a mechanism, involving the match/mismatch hypothesis², by which variability in temperature affects larval cod survival and conclude that rising temperature since the mid-1980s has modified the plankton ecosystem in a way that reduces the survival of young cod.

Fish stock biomass fluctuates in space and time, but the causes

and mechanisms responsible for the observed variability remain poorly identified³. A number of studies have reported that temperature influences cod recruitment, although the relationships found were often weak for the North Sea⁴. Food quantity and quality are also essential because they influence the growth of fish larvae^{2,5} and attempts have been made to establish relationships between changes in plankton (that is, availability of prey) and cod⁶. However, such studies, including the ICES/GLOBEC Cod and Climate Change programme⁷, often focused on a limited number of biological parameters and no credible evidence has emerged^{4,8}.

Here we have simultaneously used six key biological parameters for the diet and growth of cod larvae and juveniles in the North Sea^{5,9}. We considered the total calanoid copepod biomass as a quantitative indicator of food for larval cod, the mean size of calanoid copepods as a qualitative indicator of food, and the abundance (that is, mean number of individuals per sample) of the two dominant congeneric species, *Calanus finmarchicus* and *C. helgolandicus*, of the genus *Pseudocalanus* and of the taxonomic group euphausiids. These indicators were derived using data (46,777 samples) collected by the Continuous Plankton Recorder (CPR) survey¹⁰. We first examined long-term monthly changes (period 1958–1999) in the plankton ecosystem of the North Sea (including the Skagerrak), applying a standardized Principal Component Analysis (PCA, table years–months × indicators). The first principal component (Fig. 1, 33.71% of the total variability) revealed a clear distinction between the periods 1963–1983 and both of the periods 1984–1999 and 1958–1962. The period 1963–1983 was characterized by high abundance of prey for larval cod (positive anomalies in the biomass of calanoid copepods, in the abundance of *C. finmarchicus*, euphausiids and *Pseudocalanus* spp.) and a high mean size of calanoid copepods. In twelve of the 21 years from 1963 to 1983 cod recruitment (one-year-olds) in the North Sea was high, in parallel with positive anomalies in the plankton ecosystem (Fig. 1). Estimated recruitment during this period was

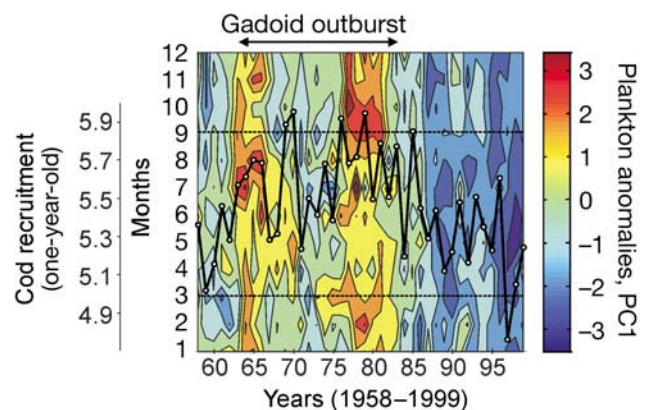


Figure 1 Long-term monthly changes (1958–1999) in the plankton index (as the first principal component, 33.78% of the total variability), resulting from analysis of the table years–months × biological indicators. The main variables related to this first principal component were, in order of importance, mean abundance (as mean number of individuals per CPR sample) of *C. finmarchicus* (normalized first eigenvector $C_m = 0.84$), euphausiids ($C_m = 0.72$), mean size of calanoid copepod ($C_m = 0.72$), *C. helgolandicus* ($C_m = -0.41$), calanoid copepod biomass ($C_m = 0.34$) and the genus *Pseudo-calanus* spp. ($C_m = 0.07$). A negative anomaly in the first principal component indicates a low value for all biological parameters with the exception of *C. helgolandicus* (opposite pattern) and *Pseudocalanus* spp. (no relationship). Cod recruitment (one-year-olds; in decimal logarithm) in the North Sea (curve in black) is superimposed with a lag of one year. The period of the 'gadoid outburst'¹¹ is indicated. Horizontal dashed lines indicate the period (March–September) of larval cod occurrence in the North Sea.

higher than at any time since 1921 (ref. 3) and the resultant high biomass of cod has been called the 'gadoid outburst'¹¹. Cod recruitment decreased from the mid-1980s, coincident with unfavourable changes in the plankton ecosystem, compared to the earlier period 1963–1983.

To investigate further the relationships between plankton and cod recruitment, we performed a standardized PCA on the six plankton indicators for each month from March to September for the period 1958–1999 (42 table yr $\times$ 7 months = 6 indicators). The 7-month period corresponded to the main period of larval/juvenile cod occurrence in the North Sea², a time when they experience high mortality¹², and the period of maximum changes in plankton (see Fig. 1). Long-term changes in cod recruitment (one-year-olds) significantly covaried positively with changes in plankton (as represented by the first principal component: PC1; 27.87% of the total variability) one year earlier (Fig. 2a, b). When the long-term variability was removed, the relationship between both variables was still highly significant (Fig. 2c). Because cod recruitment depends on the reproductive output of the spawning fish, we fitted a Ricker-type¹³ stock-recruitment relationship for the years 1946–1999 and took the residuals from that relationship as an index of the survival in each year. This index of larval cod survival was also significantly positively correlated with the plankton index. The correlations indicate that changes in plankton affect survival of cod larvae/juveniles, resulting in bottom-up control. Our results explain the increase in recruitment and therefore cod biomass during the 'gadoid outburst' as a consequence of a plankton ecosystem that was highly favourable for the survival of larval/juvenile cod. This increase occurred despite the increase in fishing mortality during this period^{3,14}.

The match/mismatch hypothesis relates survival to the match between the time of larval occurrence and that of the production of their food². The critical period for this match probably extends through the late-larval/early-juvenile stages and the prey must be of

a suitable size^{15–17}. Our results are consistent with this hypothesis and with a recent study that relates survival of larval haddock to the timing of the phytoplankton bloom¹⁸. For copepods, our results suggest that the link found with cod is primarily due to qualitative changes. Examination of variables contributing to the first principal component of plankton (see Fig. 2) showed that the relationship between PC1 and cod was mainly due to changes in the mean size of calanoid copepods (as indicated by the normalized eigenvectors, the mean contribution of this variable to the first principal component for the 7-month period was on average $C_{7m} = 46.24\%$). The mean size of calanoid copepods decreased by a factor of two after the beginning of the 1980s (Fig. 3a). The ratio between length of prey and cod, which is thought to remain constant throughout the larval and early juvenile stages (about 0.05 according to Munk's model⁵), was calculated for cod from 15 to 45 mm in length and for adult calanoid copepods in July. We found a mismatch between the size of prey and cod for fish larger than 30 mm after the mid-1980s (ratio less than 0.03 according to Munk's model⁵, Fig. 3b). The diminution in the mean size of calanoid copepods is a probable consequence of major changes in the community structure of calanoid copepods recently detected in the northeast Atlantic and attributed to increasing temperature^{19,20}.

Changes in the abundance of *C. finmarchicus* ($C_{7m} = 44.89\%$) and *C. helgolandicus* ($C_{7m} = 23.04\%$) also contributed to PC1. *Calanus* (from eggs to adults) is an important prey for cod larvae/juveniles until July–August^{5,9}. The progressive substitution of *C. finmarchicus* (Fig. 3c) by *C. helgolandicus* (Fig. 3d) has delayed the timing of occurrence of *Calanus* prey from spring to late summer at a time when larval/juvenile cod feed more on euphausiids and other fish larvae⁹. Although the total biomass of copepods ($C_{7m} = 3.61\%$) and the abundance of the genus *Pseudocalanus* ($C_{7m} = 0\%$) did not contribute to PC1, examination of their long-term monthly changes also shows a decrease after the beginning of the 1980s (Fig. 3e). A long-term decrease in euphausiids (Fig. 3f)

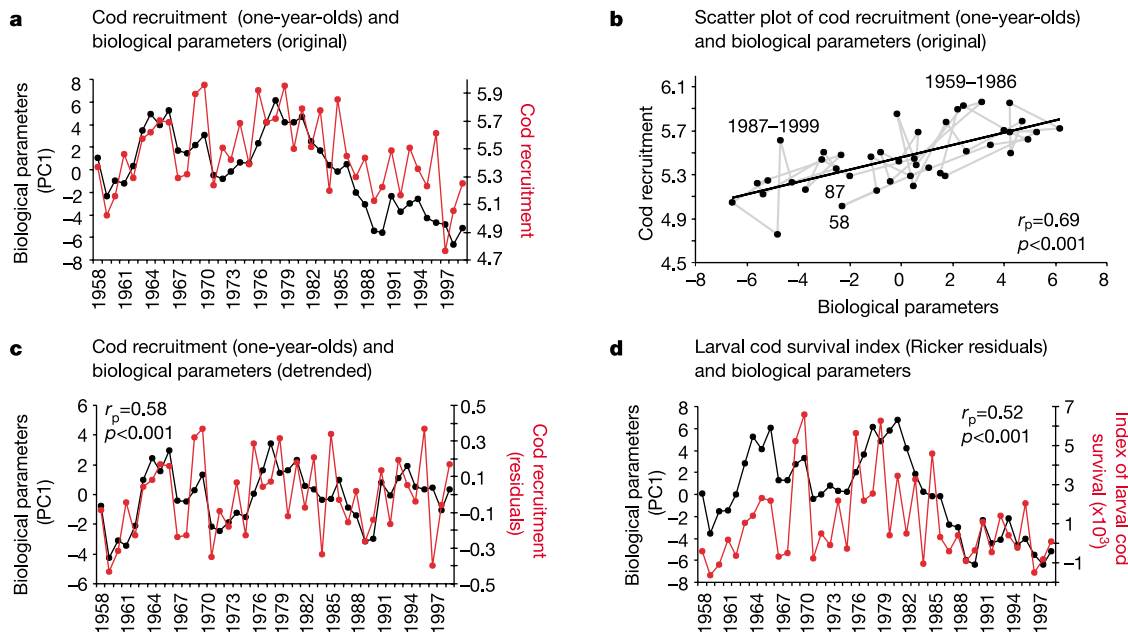


Figure 2 Relationships between plankton fluctuations and cod recruitment (one-year-olds) at lag one. **a**, Long-term changes (1958–1999) in the plankton index (as the first principal component, in black, explaining 27.87% of the total variability), resulting from analysis of the table years $\times$ months–indicators and long-term changes (1959–2000) in cod recruitment (in red, in decimal logarithm). **b**, Scatter plot of cod recruitment (at lag

one) versus the plankton index. Key years and periods are indicated. **c**, Detrended plankton index (in black) and cod recruitment (in red, in decimal logarithm). **d**, Long-term changes in both the larval cod survival and plankton indices. The coefficient of linear correlation (r_p) and its associated probability (p) are indicated.

was significantly related to PC1 ($C_{7m} = 31.36\%$). Euphausiids also constitute an important part of the diet of cod larvae and juveniles⁹. In particular, euphausiids have a high-energy content and constitute an important source of vitamin A for fish such as cod, which cannot synthesize this vitamin²¹. Thus, all biological parameters, important for the diet and growth of larval/juvenile cod, contribute to explaining the long-term changes in cod recruitment. The major changes during the mid-1980s corroborate the hypothesis of a regime shift in the North Sea²². This switch radically changed the plankton environment of larval/juvenile cod. Since then, it has been unfavourable for the survival of young cod.

Plankton fluctuations (as PC1) are significantly correlated to sea surface temperature changes in the North Sea (original data: $r = -0.63$, $p < 0.001$; detrended data: $r = -0.58$, $p < 0.001$).

Temperature covaried less with cod recruitment (original data: $r = -0.56$, $p < 0.001$; detrended data: $r = -0.50$, $p < 0.001$). Increasing temperature may have had a doubly negative impact on cod survival in the North Sea. First, temperature increases cod metabolism and energetic cost²³; thus, when food is limited, the optimal temperature for growth of cod decreases²⁴. Second, increasing temperature reduces the number of prey available for larval/juvenile cod. The resultant diminution in growth may reduce survival (the hypotheses of size-specific survival²⁵ or growth-dependent mortality²⁶) and lead to poor recruitment. Our results, therefore, provide evidence that unfavourable changes in the plankton ecosystem have exacerbated the impact of overfishing¹ in reducing recruitment of North Sea cod since the mid-1980s. Changes in the plankton ecosystem are also the probable cause of

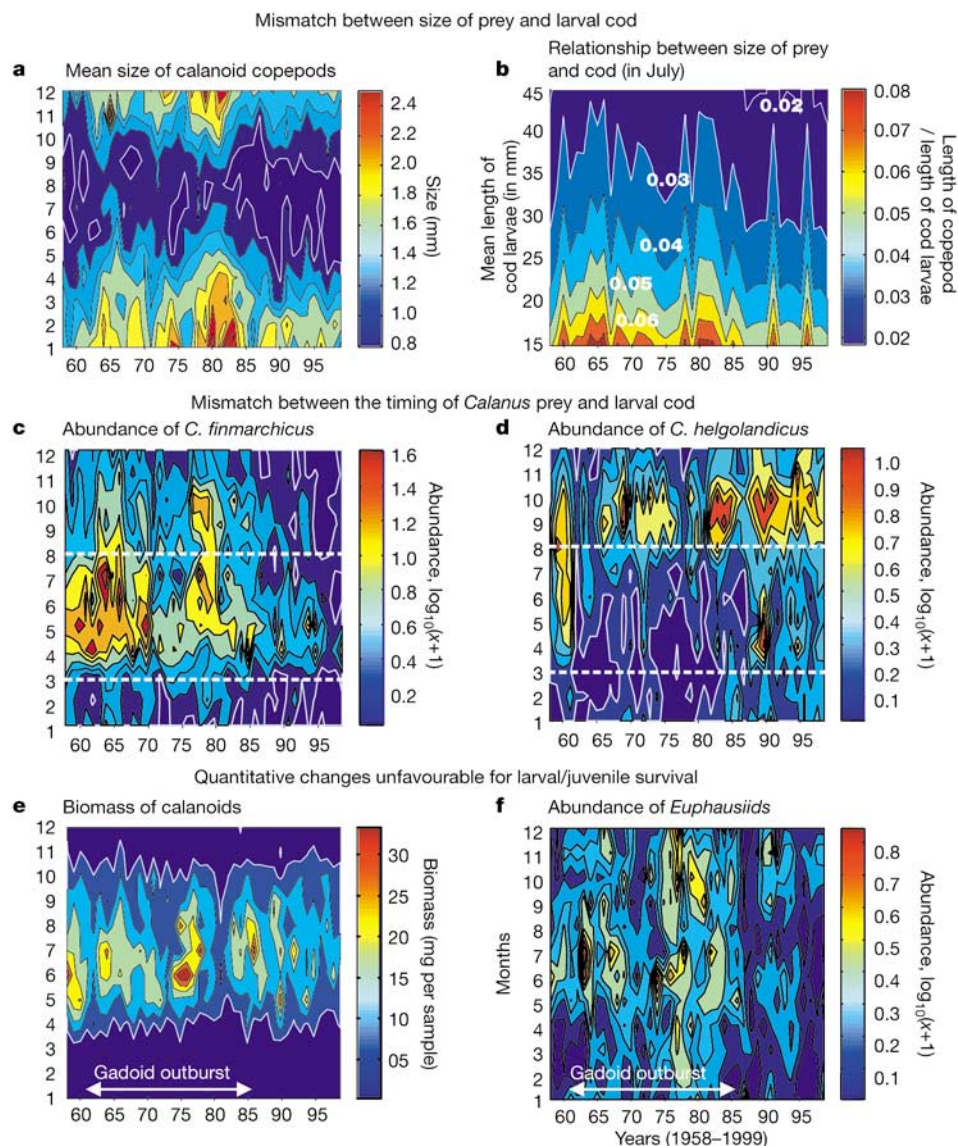


Figure 3 Plankton changes (1958–1999) and their consequences for larval/juvenile cod survival. **a**, Long-term monthly changes in the mean size of calanoid copepods (minimum size of female). **b**, Long-term changes in the ratio size of calanoid copepod:length of cod larvae. Larval cod (from 15 to 45 mm long) feed on adult *C. finmarchicus* in July⁹ and according to Munk's model⁵, the ratio should be close to 0.05, becoming highly unfavourable below 0.03 and above 0.08. **c**, Long-term monthly changes in the abundance (as mean number of individuals per CPR sample) of *C. finmarchicus*, showing

a decrease after the mid-1980s. **d**, Long-term monthly changes in the abundance (as mean number of individuals per CPR sample) of *C. helgolandicus*. The period in which cod feeds on *Calanus* is indicated by the dashed grey horizontal lines. **e**, Long-term monthly changes in the total biomass of calanoid copepods. **f**, Long-term monthly changes in the abundance (as mean number of individuals per CPR sample) of euphausiids. The period of the 'gadoid outburst' is indicated.

the increased recruitment during the period 1963–1983, which resulted in the ‘gadoid outburst’.

Methods

Plankton data

Biological data used in this study were collected by the CPR survey. This is an upper-layer plankton monitoring programme that has been operated on a routine monthly basis in the North Atlantic and North Sea since 1946¹⁰. Sampling is carried out by a high-speed plankton recorder (about 20 km h⁻¹) that is towed behind merchant ships at monthly intervals on regular routes at a standard depth of approximately 6.5 m. One CPR sample corresponds to about 3 m³ of sea water filtered¹⁰.

Selection of plankton indicators

In the North Sea, cod spawn in March and eggs start to hatch a few weeks later². From March to September, feeding of cod larvae/juveniles gradually progresses from mainly copepod eggs (April) to copepod and euphausiid nauplii (May), then a copepod-dominated diet until July and finally a progressive replacement of the copepod-based diet by euphausiids and other fish larvae from August⁹. Among copepods, *C. finmarchicus* is by far the dominant species eaten by larval cod, followed by *Pseudocalanus* spp⁵. Therefore, the abundance (number of individuals per CPR sample) of *Calanus finmarchicus*, its congeneric species *C. helgolandicus*, *Pseudocalanus* spp. and euphausiids was assessed in the North Sea (including the Skaggeirak). The size composition of prey is also a crucial parameter². The mean size of calanoid copepod (minimum size of female) per CPR sample was also calculated. We chose the minimum size of female because adult females, or copepodite stage V, represent the majority of copepods caught in the samples²⁷. Total calanoid copepod biomass per CPR sample was used as a quantitative indicator of food for larval/juvenile cod and was estimated from the size of each calanoid copepod (a total of 108 calanoid species), their abundance and allometric relationships²⁸.

Cod data

Cod data on recruitment (one-year-olds) were derived from virtual population analysis²⁹. Data for 1963–2000 are from ref. 14 and for 1958–1962 from ref. 3. These two sources overlap for the 31-yr period 1963–1993. The recruitment values taken from ref. 3 were adjusted using a linear regression analysis for the 31-yr period of overlap to produce a time series of recruitment from 1958.

Correlation analysis

The Pearson linear correlation was calculated between cod recruitment (one-year-olds) at lag one and the plankton index on original and detrended time series to take into account temporal autocorrelation. Series were detrended by the use of Singular Spectrum Analysis³⁰. The method uses a principal component analysis performed on an autocovariance matrix (also called a Toeplitz matrix) to decompose a time series into a succession of signals of decreasing variance. The procedure is fully described in ref.20. The long-term trend of each time series was first assessed by using both the eigenvectors and the principal components representing the low-frequency variability. Then, the detrended time series were calculated by subtracting the original time series by their respective long-term trend.

Received 12 August; accepted 28 October 2003; doi:10.1038/nature02164.

1. Cook, R. M., Sinclair, A. & Stefansson, G. Potential collapse of North Sea cod stocks. *Nature* **385**, 521–522 (1997).
2. Cushing, D. H. Plankton production and year-class strength in fish populations: an update of the match/mismatch hypothesis. *Adv. Mar. Biol.* **26**, 249–292 (1990).
3. Pope, J. G. & Macer, C. T. An evaluation of the stock structure of North Sea cod, haddock, and whiting since 1920, together with a consideration of the impacts of fisheries and predation effects on their biomass and recruitment. *ICES J. Mar. Sci.* **53**, 1157–1169 (1996).
4. Brander, K. M. Effects of environmental variability on growth and recruitment in cod (*Gadus morhua*) using a comparative approach. *Oceanol. Acta* **23**, 485–496 (2000).
5. Munk, P. Prey size spectra and prey availability of larval and small juvenile cod. *J. Fish Biol.* **51** (suppl. A), 340–351 (1997).
6. Rothschild, B. J. Year class strengths of zooplankton in the North Sea and their relation to cod and herring abundance. *J. Plankton Res.* **20**, 1721–1741 (1998).
7. ICES Cooperative Research Report *Report of the Workshop on Gadoid Stocks in the North Sea during the 1960s and the 1970s. 4th ICES/GLOBEC Backward-facing Workshop* (eds Heath, M. R. & Brander, K. M.) (ICES, Copenhagen, 2001).
8. Myers, R. A., Hutchings, J. A. & Barrowman, N. J. Hypotheses for the decline of cod in the North Atlantic. *Mar. Ecol. Prog. Ser.* **138**, 293–308 (1996).
9. Thorisson, K. The food of larvae and pelagic juveniles of cod (*Gadus morhua* L.) in the coastal waters west of Iceland. *Rapp. P.-v. Reun. Cons. Perm. Int. Explor. Mer.* **191**, 264–272 (1989).
10. Warner, A. J. & Hays, G. C. Sampling by the continuous plankton recorder survey. *Prog. Oceanogr.* **34**, 237–256 (1994).
11. Cushing, D. H. The gadoid outburst in the North Sea. *J. Cons. Cons. Int. Explor. Mer.* **41**, 159–166 (1984).
12. Letcher, B. H., Rice, J. A., Crowder, J. A. & Rose, K. A. Variability in survival of larval fish: disentangling components with a generalized individual based model. *Can. J. Fish. Aquat. Sci.* **53**, 787–801 (1996).
13. Ricker, W. E. Stock and recruitment. *J. Fisher. Res. Bd Can.* **11**, 559–623 (1954).
14. ICES Cooperative Research Report. *Report of the ICES Advisory Committee on Fish Management* (ICES, Copenhagen, 2002).
15. Bollens, S. M. et al. Seasonal plankton cycles in a temperate fjord and comments on the match-mismatch hypothesis. *J. Plankton Res.* **14**, 1279–1305 (1992).
16. Gotcheitas, V., Puvanendran, V., Leader, L. L. & Brown, J. A. An experimental investigation of the

- match/mismatch hypothesis using larval Atlantic cod. *Mar. Ecol. Prog. Ser.* **130**, 29–37 (1996).
17. Cushing, D. H. *Towards a Science of Recruitment in Fish Populations* (Ecology Institute, Oldendorf/Luhe, 1997).
18. Platt, T., Fuentes-Yaco, C. & Frank, K. T. Spring algal bloom and larval fish survival. *Nature* **423**, 398–399 (2003).
19. Beaugrand, G., Reid, P. C., Ibañez, F., Lindley, J. A. & Edwards, M. Reorganisation of North Atlantic marine copepod biodiversity and climate. *Science* **296**, 1692–1694 (2002).
20. Beaugrand, G. & Reid, P. C. Long-term changes in phytoplankton, zooplankton and salmon linked to climate. *Glob. Change Biol.* **9**, 801–817 (2003).
21. Mauchline, J. & Fisher, L. R. The biology of euphausiids. *Adv. Mar. Biol.* **7**, 1–454 (1969).
22. Reid, P. C., Borges, M. & Svendsen, E. A regime shift in the North Sea circa 1988 linked to changes in the North Sea horse mackerel fishery. *Fish. Res.* **50**, 163–171 (2001).
23. Otterlei, E., Nyhammer, G., Folkword, A. & Stefansson, S. O. Temperature- and size-dependent growth of larval and early juvenile Atlantic cod (*Gadus morhua*): a comparative study of Norwegian coastal cod and northeast Arctic cod. *Can. J. Fish. Aquat. Sci.* **56**, 2099–2111 (1999).
24. Swain, D. P. et al. Density- versus temperature-dependent growth of Atlantic cod (*Gadus morhua*) in the Gulf of St. Lawrence and on the Scotian Shelf. *Fish. Res.* **59**, 327–341 (2003).
25. Pepin, P. An appraisal of the size-dependent mortality hypothesis for larval fish: comparison of a multispecies study with an empirical review. *Can. J. Fish. Aquat. Sci.* **50**, 2166–2174 (1993).
26. Gallego, A. & Heath, M. The effect of growth-dependent mortality, external environment and internal dynamics on larval fish otolith growth: an individual-based modelling approach. *J. Fish Biol.* **51** (suppl. A), 121–134 (1997).
27. Mauchline, J. *The Biology of Calanoid Copepods* (Academic, San Diego, 1998).
28. Peters, H. *The Biological Implication of Body Size* (Cambridge Univ., Cambridge, 1983).
29. Shepherd, J. G. Extended survivors analysis: an improved method for the analysis of catch-at-age data and abundance indices. *ICES J. Mar. Sci.* **56**, 584–591 (1996).
30. Vautard, R. P. Y. & Ghil, M. Singular-spectrum analysis: a toolkit for short, noisy chaotic signals. *Physica D* **58**, 95–126 (1992).

Acknowledgements We are grateful to all past and present members and supporters of the Sir Alister Hardy Foundation for Ocean Science whose sustained help has allowed the establishment and maintenance of the CPR data set in the long term. The main support of this work was from UK DEFRA, and both the French PNEC Art 4 and IFB ‘Biodiversité et changement global’ programmes. Consortium support for the CPR survey is provided by agencies from the following countries: UK, USA, Canada, the Faroe Islands, France, Ireland, the Netherlands, Portugal, the IOC and the European Union. We thank the owners, masters and crews of the ships that tow the CPRs on a voluntary basis.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.B. (gbea@mail.pml.ac.uk).

Visual control of action but not perception requires analytical processing of object shape

Tzvi Ganel & Melvyn A. Goodale

CIHR Group on Action and Perception, Department of Psychology, University of Western Ontario, London, Ontario N6A 5C2, Canada

The visual perception of object shape depends on ‘holistic’^{1–4} processing in which a given dimension cannot be perceptually isolated from the other dimensions of the object. The visual control of action (such as grasping an object), however, which is mediated by cortical areas that are largely independent of those mediating conscious perception^{5–8}, must take into account only the most action-relevant dimension of an object without being misled by other non-relevant object features. Here we report the results of two experiments showing that vision for perception and vision for action deal with objects in a fundamentally different manner. We tested participants’ ability to make perceptual judgements of the width of different rectangular objects or to grasp them across their width, while in both cases ignoring length^{9,10}. Participants could not ignore length when making perceptual judgements of width but they could completely ignore length when grasping the same objects. These results suggest that in situations in which the elementary dimensions of an object’s

shape are perceived in a holistic manner, the same dimensions are treated analytically when a visually guided action is directed at that same object.

The idea that vision treats object shape in a holistic manner has been a basic theme running through theoretical accounts of perception from early gestalt psychology¹¹ to more contemporary cognitive science^{1,12–14}. Indeed, encoding an object holistically permits a representation of the object that preserves the relations between object parts and its surroundings without requiring precise information about the absolute size of the object's dimensions^{15,16}. When we interact with an object, however, it is imperative that the visual processes controlling the action take into account the absolute metrics of the most relevant dimension of the object without being influenced by other dimensions or features¹⁷. In other words, rather than being holistic, the visual processing mediating action should be analytical.

To test this idea, which on the face of it might seem counter-intuitive, we carried out two experiments in which we compared how visually guided action and visual perception dealt with one of the fundamental attributes of an object, its shape^{18–21}. In both experiments we used a modified version of a well-established psychophysical tool, Garner's speeded-classification task^{9,10,22,23}, which provides a reliable measure of how efficiently people can process one dimension of an object while ignoring its other dimensions²⁴. In a typical Garner experiment participants are asked to classify objects on the basis of a single dimension under two different conditions. In one condition (baseline), the relevant dimension varies but another, irrelevant, dimension is kept constant. In the other condition (filtering), the relevant dimension again varies but this time so does the irrelevant dimension. If participants are able to process the two dimensions independently (that is analytically), then the speed and accuracy should be identical for the baseline and filtering conditions. If participants cannot process the two dimensions independently and must treat them holistically, then performance should be worse for the filtering condition than for the baseline condition, because participants cannot 'filter out' the changes in the irrelevant dimension.

The dimensions that we used were the width and length of rectangular objects. We chose these dimensions for three reasons. First, they represent an elementary instance of dimensions that constitute the shape of objects¹⁸. Second, these dimensions have been shown to be represented holistically in a number of perceptual studies^{25–27}. Finally, the width and length of objects can be manipulated in a modified Garner's task to compare directly the processing underlying perception and visually guided action within the same experiment.

Four different rectangular objects were used in both of our experiments. They were created from factorial combination of

two different widths with two different lengths. The two longer objects were presented in random order in one block of the baseline condition and the two shorter ones in a second block of the baseline condition. All four objects were presented in the filtering condition. In all cases, width was the only relevant dimension.

In experiment one we used Garner's paradigm to compare object perception and object-directed action. In the perceptual classification task 12 participants made speeded judgements (wide or narrow) of the width of each rectangular object by pressing one of two response buttons (Fig. 1). In the grasping task, the objects were presented in the same way but now the participants were instructed to reach out and grasp each object as quickly as possible across its width using their finger and thumb in a natural precision grip. Hand and finger movements were tracked using opto-electronic recording of the position of infrared light-emitting diodes attached to index finger, thumb and wrist.

Mean reaction times for each participant were calculated for correct responses on perceptual trials and successful final pick-ups on grasping trials during baseline and filtering conditions. Outliers of more than two standard deviations above the mean were eliminated. Accuracy for both perceptual classification and grasping was high (less than 1% error) and did not vary between conditions.

As can be seen in Fig. 2, the reaction times for perceptual classification in the filtering condition (436 ± 20.4 ms; mean $\pm$ standard error) were 23 ms slower than reaction times in the baseline condition (413 ± 25 ms). This significantly worse performance in the filtering compared with the baseline condition ($t_{11} = 2.49$, $P < 0.05$), sometimes called Garner interference, indicates that the width of the objects could not be processed independently of length when perception was required. In contrast, however, reaction times to complete the movement in the grasping task in the filtering condition (856 ± 21.5 ms) were similar to those in the baseline condition (859 ± 22.3 ms). The -3 ms difference between the conditions was not significant ($t_{11} = 0.3$, $P > 0.1$). The same absence of a difference between filtering and baseline conditions was evident when only the time to initiate the grasping movement was considered or the time to reach maximum grip aperture. An analysis of variance (ANOVA) between tasks (reaction times for perceptual judgements versus reaction times to grasp objects) and conditions (baseline versus filtering) revealed a two-way interaction between these factors, confirming that width and length of rectangles are treated differently for perception and visually guided action ($F_{1,11} = 7.93$, mean square error = 251, $P < 0.05$).

To test whether grip scaling was sensitive to width and to test for possible differences in grip scaling between baseline and filtering conditions, we calculated the maximum grip aperture for wide and narrow objects in each experimental condition. The maximum

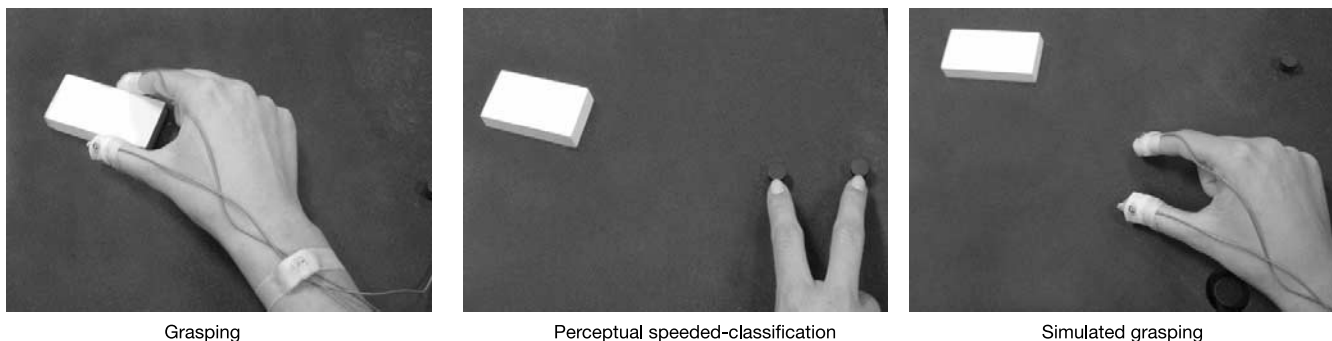


Figure 1 Tasks used in experiments one and two. From left to right, the tasks shown are: the grasping task (left) and the perceptual speeded-classification task (centre) (experiment one), and the simulated grasping task (right) for perceptual estimation (experiment two).

opening between the thumb and index finger during grasping, which is typically achieved 70% of the way through the movement trajectory, is known to be well correlated with the size of the goal object. Maximum grip aperture was 80.8 mm in the baseline condition and 80.2 mm in the filtering condition for wide objects, and 76 mm in the baseline condition and 75.6 mm in the filtering condition for narrow objects. An ANOVA between width and condition revealed a significant main effect of width ($F_{1,10} = 119$, mean square error = 2.06, $P < 0.001$), which indicated that grip scaling was highly sensitive to the width of an object. The main effect of condition ($F_{1,10} = 0.2$, $P > 0.1$) and the interaction between width and condition ($F_{1,10} = 0.1$, $P > 0.1$) were not significant.

The results of experiment one show that participants could successfully grasp rectangular objects across their width without being affected by irrelevant variations in length but they could not (or at least did not) make perceptual judgements of width without being affected by irrelevant variations in length. Therefore, these results show that the two elementary visual dimensions that constitute the shape of rectangles can be treated in a completely independent manner for visually guided action but not for visual perception.

An alternative explanation for the pattern of results in experiment one is that the different task demands, rather than different kinds of visual processing, led to the dissociation found between perception and visually guided action. Specifically, one has to refer to memory for what is considered wide or narrow in the experiment when making perceptual judgements, but not when grasping. A second experiment was therefore carried out to rule out this possibility. In experiment two a simulated grasping task was used to assess perceptual judgements. In this task, participants estimated the width of each object by reaching out and placing their finger and thumb a particular distance apart at a remote location without actually grasping the object (Fig. 1). This task does not require memory-based comparison and is known to make use of perceptual information²⁸. We expected, therefore, that we would find Garner interference in this task.

Eight participants took part in experiment two in a similar

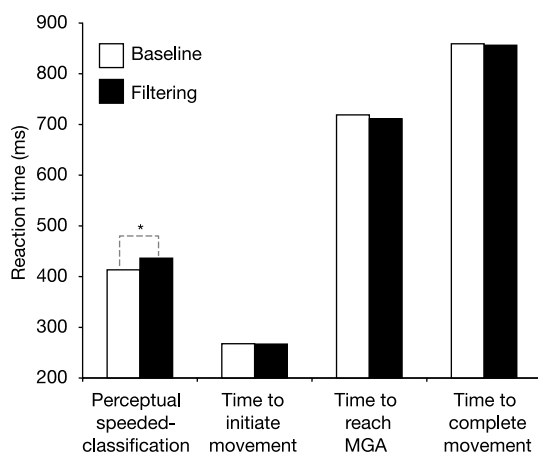


Figure 2 Effects of irrelevant variations in length on the perception of object width and on object-directed grasping (across the object's width). Reaction times for speeded-classification (first column) of width were significantly slower in filtering (filled bars) than in baseline (open bars) conditions. Reaction times for grasping (second to fourth columns) were not different between filtering and baseline conditions. Other kinematic measures of grasping also showed no difference between conditions. A control experiment that rules out the possibility that visual feedback during the execution of the grasping movement may have eliminated the Garner interference effect in this condition is described in Supplementary Information. MGA, maximum grip aperture.

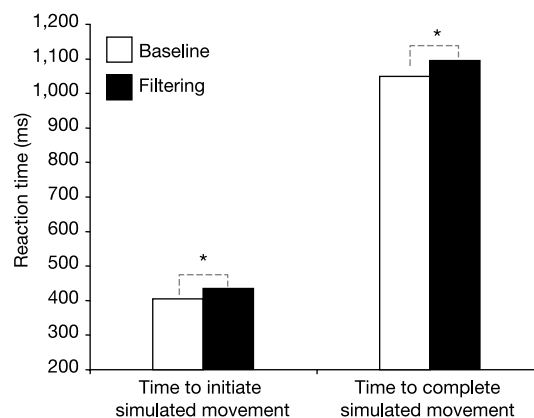


Figure 3 Effects of irrelevant variations in length on the reaction time for simulated grasping. The mean time to initiate the movement was slower in the filtering condition (filled bars) than in the baseline condition (open bars) ($t_7 = 2.39$, $P < 0.05$). The same was true for overall movement time ($t_7 = 2.40$, $P < 0.05$).

paradigm to that used in experiment one, but were asked to make simulated rather than actual grasping movements. As can be seen in Fig. 3, significant Garner interference effects were found in this experiment for both the time to initiate and the time to complete the movements. These results confirmed our prediction that width and length are perceived holistically even when memory-based comparisons are not required. Notably, our design also allowed us to examine the effects of length on computations of width using another measure. Specifically, if width cannot be perceived independently of length, then one would expect an illusion to occur, in which longer objects are perceived to be narrower than shorter objects with the same width. This was exactly what we found for the width estimates in experiment two (Fig. 4). This was not the case, however, for maximum grip aperture in the grasping task of experiment one; in this case, there was no effect of the object's length on computation of width. Taken together, the findings of

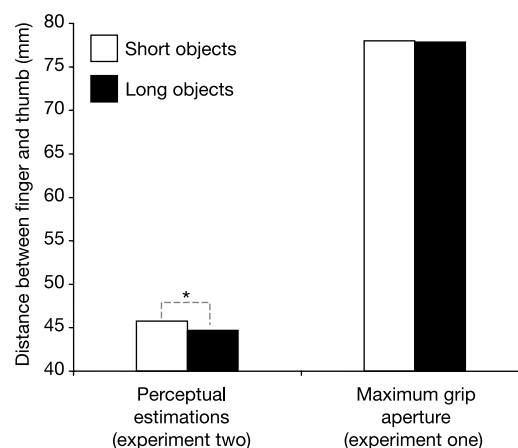


Figure 4 Effects of length on grip scaling for width in simulated grasping and real grasping in the filtering condition. In experiment two, in which perceptual estimates were required for the simulated response, longer objects (filled bars) resulted in significantly smaller finger–thumb apertures than shorter objects (open bars) ($t_7 = 3.01$, $P < 0.05$). In experiment one there was no difference in maximum grip aperture for longer and shorter objects ($t_7 = 0.65$, $P > 0.1$). A significant interaction between length (longer objects, shorter objects, within-subject variable) and task (simulated grasping, real grasping, between-subjects variable) showed that the difference between these effects was significant ($F_{1,17} = 4.86$, mean square error = 0.39, $P < 0.05$).

experiments one and two provide compelling support for the idea that perception of an object's width is affected by its length but grasping across the object's width is not.

The fact that the time taken to initiate and complete a grasping movement is not affected by variations in the irrelevant dimension supports the notion that the visual information used to program the grasp is not derived from the visual processing mediating perception. Right from the start, in other words, grasping selectively processes information about the relevant dimension without first processing the entire shape of the object.

The idea that visual perception uses holistic processing of object features whereas the visual control of action is more analytical helps to explain why separate cortical pathways have evolved for these two different kinds of visual processing: a ventral stream for perception and a dorsal stream for action^{5–8}. It also helps to explain why visual illusions (which, by definition, have strong effects on perceptual judgements) often have little or no effect on the scaling of grasping movements²⁹.

Most contemporary theories of visual perception (see ref. 20 for a review) agree that, for proper recognition of objects to occur, the raw visual input must be transformed into visual representations that preserve the relative aspects of an object's dimensions, as well as their relation to the dimensions, orientation and location of other objects in the scene. An obvious cost of relying on such allocentric¹⁷ and holistic representations is that important attributes of our visual environment, such as the absolute metrics of objects, are less accessible. Our findings suggest that the absolute metrics of action-relevant dimensions may be separately computed by the system mediating our interactions with objects in our near environment. Our findings also suggest that this is accomplished because, unlike with visual perception, the visual mechanisms mediating action are able to process the most relevant dimension while at the same time ignoring changes in the other, irrelevant, dimensions. In short, vision for action operates in an analytical rather than a holistic fashion. □

Methods

Experimental design

Participants were seated in front of a black tabletop on which the objects were placed at a viewing distance of approximately 40 cm. Computer-controlled PLATO goggles (Translucent Technologies) with liquid-crystal shutter lenses were used to control stimulus exposure time. Grip scaling and reaction times were recorded by an Optotrak (Northern Digital), which tracked the three-dimensional position of three infrared light-emitting diodes attached separately to the participant's index finger, thumb and wrist with small pieces of surgical tape, which allowed complete freedom of movement of the hand and fingers. Participants were right-handed undergraduate students and received a course credit for their participation. Each participant gave informed consent and the experiments were approved by the local ethics committee. To prevent participants from adopting an automated grasping strategy, the target objects were presented in random locations on the tabletop (within a circle of 5 cm) at randomly interleaved orientations between 15–25° from a plane parallel with the participant's midline (Fig. 1).

Object design and presentation

The objects were created from a factorial combination of two different widths (that is, 35.7 and 30 mm) by two different lengths (that is, 75 and 63 mm). The proportions used were the same as those used in the original perceptual study that applied Garner's task to examine the relationship between width and length²⁵. The thickness of all stimuli was 15 mm. Each object was presented 16 times (2,000 ms for each object) in the baseline condition and 16 times in the filtering condition. To equate the overall number of objects in the baseline and filtering conditions, the filtering condition was divided into two equal smaller filtering blocks (32 stimuli presentations in each block). Before the beginning of each block participants were shown the objects that would be presented in that specific block. The interstimulus interval was 3,000 ms. Four practice trials were administered before the beginning of each block and there were 1-min rests between blocks.

Experiment one

For the perceptual speeded-classification task in experiment one, participants were asked to place the first and second fingers of their right hand on the response box, and immediately after the opening of the goggles, to press the right key if the object was wide, or the left key if the object was narrow (Fig. 1). To ensure that they received the same tactile feedback in the perception task as they did in the grasping task, participants were asked to reach out and grasp each object by its width immediately after making each speeded-classification response. In the grasping task of experiment one, participants placed the

index finger and thumb of their right hand on a start button, and were asked to reach out and grasp each object across its width as quickly as possible immediately after the goggles were opened. Order of the conditions (baseline/filtering) and task presentation (speeded classifications/grasping) was counterbalanced across participants. Grip scaling data of one participant in experiment one was excluded from the analysis due to the loss of more than 50% of data point signals.

Experiment two

For the simulated grasping in experiment two the same apparatus was used, but now participants were asked to simulate the grasping movement that they would make if they were to pick up each object across its width. Instead of grasping the object, however, they indicated how they would do so by reaching out and placing their index finger and thumb beside the object as though they were picking it up. To ensure that they received the same tactile feedback in simulated grasping as in real grasping, participants were asked to reach out and grasp each object immediately after completing the simulated grasping movement.

Received 20 June; accepted 9 October 2003; doi:10.1038/nature02156.

- Hochstein, S. & Ahissar, M. View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron* **36**, 791–804 (2002).
- Natsopoulos, T. What are the objects of perceptual consciousness? *Am. J. Psychol.* **96**, 435–467 (1983).
- Pomerantz, J. R. & Pristach, E. A. Emergent features, attention, and perceptual glue in visual form perception. *J. Exp. Psychol. Hum. Percept. Perform.* **15**, 635–649 (1989).
- Navon, D. Forest before trees: The precedence of global features in visual perception. *Cogn. Psychol.* **9**, 353–383 (1977).
- Goodale, M. A. & Milner, A. D. Separate visual pathways for perception and action. *Trends Neurosci.* **15**, 20–25 (1992).
- Goodale, M. A. *et al.* Separate neural pathways for the visual analysis of object shape in perception and prehension. *Curr. Biol.* **4**, 604–610 (1994).
- James, T. W., Humphrey, G. K., Gati, J. S., Menon, R. S. & Goodale, M. A. Differential effects of viewpoint on object-driven activation in dorsal and ventral streams. *Neuron* **35**, 793–801 (2002).
- Humphrey, G. K., Goodale, M. A., Jakobson, L. S. & Servos, P. The role of surface information in object recognition: studies of a visual form agnostic and normal subjects. *Perception* **23**, 1457–1481 (1994).
- Garner, W. R. *The Processing of Information and Structure* (Potomac, Erlbaum, 1974).
- Garner, W. R. Selective attention to attributes and to stimuli. *J. Exp. Psychol. Gen.* **107**, 287–308 (1978).
- Koffka, K. *Principles of Gestalt Psychology* (Harcourt Brace, New York, 1935).
- Law, M. B. & Abrams, R. A. Object-based selection within and beyond the focus of spatial attention. *Percept. Psychophys.* **64**, 1017–1027 (2003).
- Duncan, J. Selective attention and the organization of visual information. *J. Exp. Psychol. Gen.* **113**, 501–517 (1984).
- O'Craven, K. M., Downing, P. E. & Kanwisher, N. fMRI evidence for objects as the units of attentional selection. *Nature* **401**, 584–587 (1999).
- Goodale, M. A. & Servos, P. in *Advances in Motor Learning and Control* (ed. Zelaznik, H. N.) 87–121 (Human Kinetics, Illinois, 1996).
- Swinnen, S. P., Dounskaia, N. & Duysens, J. Patterns of bimanual interference reveal movement encoding within a radial egocentric reference frame. *J. Cogn. Neurosci.* **14**, 463–471 (2001).
- Goodale, M. A. & Humphrey, G. K. The objects of action and perception. *Cognition* **67**, 181–207 (1998).
- Marr, D. Visual information processing: the structure and creation of visual representations. *Phil. Trans. R. Soc. Lond. B* **290**, 199–218 (1980).
- Edelman, S. & Duvdevani-Bar, S. A model of visual recognition and categorization. *Phil. Trans. R. Soc. Lond. B* **352**, 1191–1202 (1997).
- Bar, M. Viewpoint dependency in visual object recognition does not necessarily imply viewer-centered representation. *J. Cogn. Neurosci.* **13**, 793–799 (2001).
- Poggio, T. & Edelman, S. A network that learns to recognize three-dimensional objects. *Nature* **343**, 263–266 (1990).
- Ganel, T. & Goshen-Gottstein, Y. Perceptual integrity of sex and identity of faces: further evidence for the single-route hypothesis. *J. Exp. Psychol. Hum. Percept. Perform.* **28**, 854–867 (2002).
- Shalev, L. & Algom, D. Stroop and Garner effects in and out of Posner's beam: reconciling two conceptions of selective attention. *J. Exp. Psychol. Hum. Percept. Perform.* **26**, 997–1017 (2000).
- Maddox, W. T. in *Perceptual and Decisional Separability* (ed. Ashby, F. G.) 147–180 (Erlbaum, Hillsdale, NJ, 1992).
- Felfoldy, G. L. Repetition effects in choice reaction time to multidimensional stimuli. *Percept. Psychophys.* **15**, 453–459 (1974).
- Dykes, J. R. & Cooper, R. G. An investigation of the perceptual basis of redundancy gain and orthogonal interference for integral dimensions. *Percept. Psychophys.* **23**, 36–42 (1978).
- Macmillan, N. A. & Ornstein, A. S. The mean-integral representation of rectangles. *Percept. Psychophys.* **60**, 250–262 (1998).
- Goodale, M. A., Jakobson, L. S. & Keillor, J. M. Differences in the visual control of pantomimed and natural grasping movements. *Neuropsychologia* **32**, 1159–1178 (1994).
- Kwok, R. M. & Braddick, O. J. When does the Titchener Circles illusion exert an effect on grasping? Two- and three-dimensional targets. *Neuropsychologia* **41**, 932–940 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Pomerantz, M. Morgan, S. Köhler and P. Gribble for comments on the manuscript, and G. Króliczak for his help. This research was supported by grants to M.A.G. from the Canadian Institutes of Health Research and the Canada Research Chairs Program.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.G. (tganel@uwo.ca).

Parallel colour-opponent pathways to primary visual cortex

Soumya Chatterjee & Edward M. Callaway

Systems Neurobiology Laboratories, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA, and Neuroscience Program, University of California 92093, San Diego, California, USA

The trichromatic primate retina parses the colour content of a visual scene into 'red/green' and 'blue/yellow' representations^{1–2}. Cortical circuits must combine the information encoded in these colour-opponent signals to reconstruct the full range of perceived colours³. Red/green and blue/yellow inputs are relayed by the lateral geniculate nucleus (LGN) of thalamus to primary visual cortex (V1), so understanding how cortical circuits transform these signals requires understanding how LGN inputs to V1 are organized. Here we report direct recordings from LGN afferent axons in muscimol-inactivated V1. We found that blue/yellow afferents terminated exclusively in superficial cortical layers 3B and 4A, whereas red/green afferents were encountered only in deeper cortex, in lower layer 4C. We also describe a distinct cortical target for 'blue-OFF' cells, whose afferents terminated in layer 4A and seemed patchy in organization. The more common 'blue-ON' afferents were found in 4A as well as lower layer 2/3. Chromatic information is thus conveyed to V1 by parallel, anatomically segregated colour-opponent systems, to be combined at a later stage of the colour circuit.

The retina constructs colour-opponency from cones sensitive to long (L), middle (M) and short (S) wavelengths by superimposing excitatory (ON) and inhibitory (OFF) signals from different parts of the visible spectrum^{1,2,4}. Midget ganglion cells compare photon catches between L and M cones, creating a red/green opponent axis. Small and large bistratified cells relay ON signals from S cones and OFF signals from L+M cones (blue-ON receptive fields), giving rise to a blue/yellow colour axis. And an achromatic channel is established by parasol cells, which sum their cone contributions.

These ganglion cells send their outputs to the LGN, which in turn projects at least three parallel pathways (magnocellular, parvocellular and koniocellular) to V1. Neurons in magnocellular layers carry the achromatic signal from parasol cells to layer 4C α of V1, whereas parvocellular neurons relay the red/green input of midget cells to layer 4C β ^{5–10}. Cells in koniocellular zones, which may receive blue-ON input from small bistratified cells^{7,11–13}, project to superficial layers of V1¹⁴. However, konio cells do not respect laminar borders strictly, and are often found in neighbouring parvo and magno layers. The spatial overlap of neuronal populations has confounded attempts to uniquely correlate receptive field properties with cell types and their cortical projections. Thus, it is not known whether blue-ON cells recorded in parvo layers^{3,15–18} are parvo cells projecting to layer 4C β or displaced konio cells terminating in superficial V1. It is also unclear whether konio cells carry any red/green opponency to superficial layers. Finally, some konio or parvo cells may have blue-OFF receptive fields^{3,17,19}, but their cortical targets have not been identified.

We examined the functional organization of colour-opponent input to V1 by silencing cortical neurons with the GABA_A agonist muscimol, which allowed us to record directly from LGN afferent terminals in various depths of cortex. Most afferents could be unambiguously classified into four groups: blue-ON, blue-OFF, red/green or achromatic, each with its own visual response signature. Cone-isolating, drifting sinusoidal gratings at optimal temporal and spatial frequencies were used to determine the contributions of different cone types. The cell shown in Fig. 1a

was robustly driven by S-cone-isolating gratings as well as L+M gratings, with responses shown in peristimulus time histograms (PSTHs, lower part of Fig. 1a). Furthermore, the first harmonic phases of the S and L+M responses were almost 180° apart (anti-phase); that is, the peak of the S-cone-isolating stimulus influenced the cell's firing with the same sign (ON or OFF) as the trough of the L+M grating, indicating blue/yellow colour-opponency.

We recovered the signs of these cone inputs from cone-isolating reverse-correlation experiments. Spike-triggered averages (STAs) were obtained by showing a long sequence of randomly chosen, rapidly presented stimuli (sine-wave gratings of various orientations, spatial frequencies and phases, drawn from the two-dimensional Hartley basis set²⁰), and calculating the average stimulus that elicited a spike. The time course of the S-cone-isolating STA for a group of pixels representing the cell's receptive field centre is shown in the upper part of Fig. 1a (blue trace). The horizontal axis represents the 250-ms epoch before each spike, and the vertical axis is the STA cone contrast, where +1 and –1 are the maximum

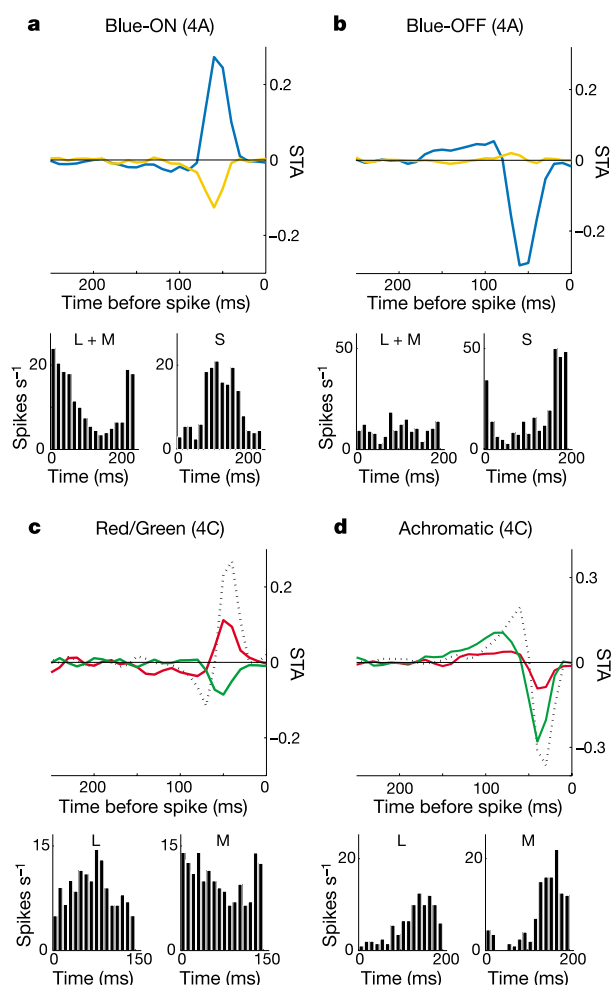


Figure 1 LGN afferents recorded in muscimol-inactivated V1 were classified into four distinct groups on the basis of cone contributions. STAs for L (M, S, L+M, L+M+S) cone-isolating stimuli are shown in red (green, blue, yellow, dotted black). **a**, Characterization of a blue-ON afferent, encountered in layer 4A of V1. Shown (lower part of the panel) are the cell's responses to cone-isolating, drifting gratings, binned over multiple grating cycles into PSTHs with horizontal axes equal to the temporal period of the stimulus. The upper part of the panel summarizes data from reverse-correlation experiments using cone-isolating Hartley stimuli. **b**, A blue-OFF afferent recorded in layer 4A, with same conventions as in **a**. **c**, A red/green afferent recording, from layer 4C. **d**, An achromatic afferent encountered in layer 4C.

and minimum achievable cone excitation for each stimulus. The upward deflection of the trace indicates an ON response to S-cone activation, whereas the downward transient seen with L+M cone-isolating stimuli (yellow trace) reveals OFF input from L and M cones, establishing this as a blue-ON cell with cone contributions very similar to those of blue-ON small bistratified cells of the retina²¹. Blue-OFF cells were characterized in the same manner (Fig. 1b). Cone-isolating gratings drove a vigorous S response and extremely weak L+M response (which was often the case for blue-OFF cells that we encountered). STAs computed for this cell confirm that the S-cone signal was OFF.

Distinguishing blue-ON and blue-OFF from red/green afferents was straightforward, as the latter were marked by L/M opponency (Fig. 1c) and little or no S-cone input. All chromatic units, in turn,

were distinct from achromatic afferents (Fig. 1d), which exhibited no opponency between cone types. (See Supplementary Fig. S1 for population data related to classification.)

To examine how chromatic inputs to V1 were organized, we reconstructed the locations of recording sites in relation to cortical layers for individual electrode penetrations, following electrolytic lesions and histology. Figure 2a shows a tangential penetration in silenced V1 traversing roughly 2.5 mm of superficial cortex from first to last recorded unit. With the electrode skirting the boundary between layer 2/3 and 4A, we found multiple clusters of blue-ON afferents (while crossing four ocular dominance transitions). Finally going through layer 4A near the end of the penetration, we encountered a cluster of blue-OFF afferents. Out of 14 penetrations with recovered lesions, 5 contained both blue-ON and blue-OFF recordings (Fig. 2a–e). Each showed this surprising input pattern, with blue-ON afferents in lower 2/3 + 4A terminating above blue-OFF. As an additional five penetrations had blue-ON with no blue-OFF (example in Fig. 2i), and given the greater number of blue-ON afferents encountered (comprising 78% of S-cone-dominated afferents in our sample), it is likely that blue-ON geniculate projections to V1 have a more extensive coverage of superficial cortex. Blue-OFF cells, on the other hand, seem to terminate in sparse patches, segregated by depth from blue-ON.

The overall laminar organization of inputs to V1 is exemplified by the penetration shown in Fig. 2e. We encountered different afferent classes in a highly specific laminar order: blue-ON → lower 2/3+4A, blue-OFF → 4A, achromatic → upper 4C, and red/green → lower 4C. This pattern was conserved across our data set. Axonal projections of magno and parvo neurons to 4C α and 4C β , respectively, are consistent with achromatic input to upper and red/green

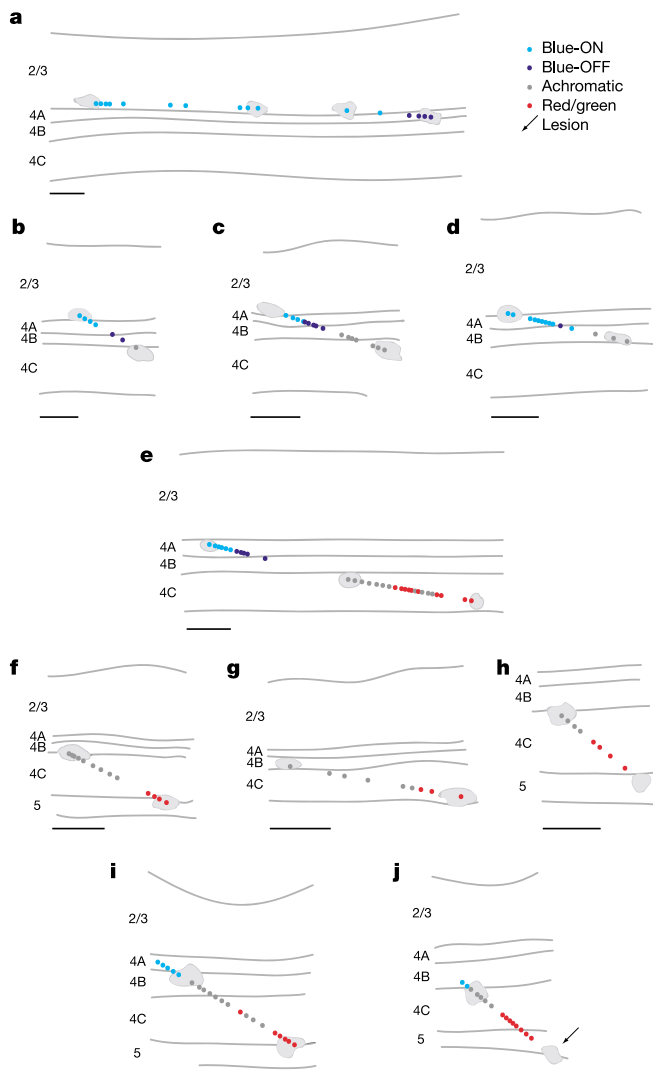


Figure 2 Reconstructions of tangential electrode penetrations in muscimol-inactivated V1. Coloured markers (light blue, dark blue, red and grey) indicate the locations of individual LGN afferent recordings and their colour opponencies (blue-ON, blue-OFF, red/green and achromatic, respectively). **a–e**, The five penetrations in which both blue-ON and blue-OFF afferents were encountered show a functional organization of S-cone-dominated afferents based on depth, with blue-ON inputs terminating above blue-OFF in almost all cases. **e–j**, Achromatic and red/green inputs were restricted to upper and lower 4C, respectively. The direction of each penetration is aligned from left to right, and lesions used for electrode track reconstructions are shown as grey patches, as indicated by the arrow in **j**. Scale bar, 250 μ m.

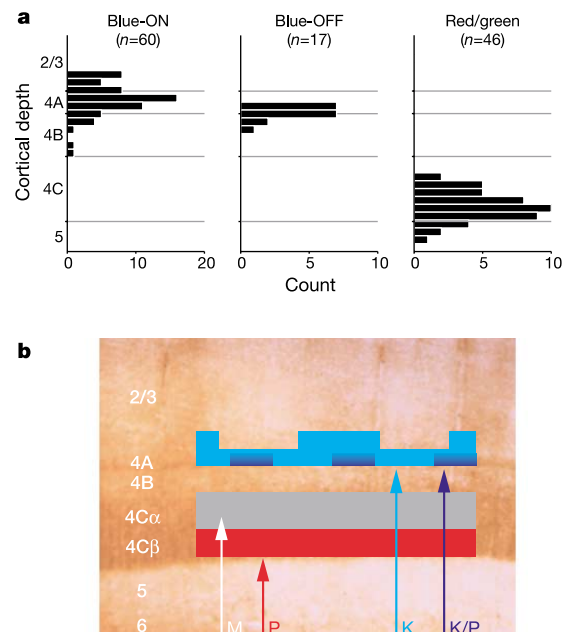


Figure 3 Laminar organization of colour-opponent afferents. **a**, Histograms show the laminar distribution of all blue-ON, blue-OFF and red/green afferents recorded in V1. Blue-OFF afferents were restricted to the lower part of the blue-ON range, and both were segregated from red/green input with no overlap. **b**, Diagram describing hypothesized relationships between physiology of colour inputs and anatomical projections. Achromatic and red/green afferents to layer 4C almost certainly arise from magno (M) and parvo (P) contributions to 4C α and 4C β , respectively. Blue-ON inputs to layers 4A and 2/3 are most probably koniocellular. Blue-OFF afferents could be koniocellular (K), but there is also evidence supporting a parvo projection to 4A, raising the possibility that a separate blue-OFF parvo substream bypasses its usual target of 4C β to project solely to 4A.

to lower 4C (Fig. 2e–j). We analysed 221 afferent recordings (60 blue-ON, 17 blue-OFF, 46 red/green, and 98 achromatic), and the laminar distribution of colour-opponent inputs shows complete segregation of blue/yellow and red/green axes (Fig. 3a). Crucially, no red/green afferents were encountered in supragranular layers, and no blue/yellow input was found in 4C.

Columnar organization of blue/yellow and red/green input is not likely. Such organization predicts some population heterogeneity at a given laminar depth as the electrode tangentially cuts across putative columns over multiple penetrations. Because no mixing of blue/yellow and red/green afferents was seen in any layer after recording from a total ~ 4.3 mm of lower 2/3 + 4A and ~ 5.4 mm of 4C (measurements post-histology, so normal tissue shrinkage of 20–40% underestimates coverage), our observations strongly imply a laminar organization instead. There is heterogeneity between blue-ON and -OFF in layer 4A, but if blue-OFF terminations are patchy, they are not columnar in the classical sense because they do not extend above or below 4A. They may, however, still be related to some overall columnar architecture of V1 (for example, blobs). Also, different classes of red/green cells (all the permutations between centre/surround and ON/OFF L- and M-cone contributions) were treated as one group in this study, because we saw no clear sublaminar or columnar organization of these classes.

Because extracellular recordings are subject to considerable sampling bias, it is possible that red/green afferents do project to upper layers and we simply did not encounter them. The same argument applies to blue/yellow input to 4C, as well as to the apparent patchiness of blue-OFF input to 4A. Also, despite focused attempts, we could not record from upper layer 2/3 or 1, where some afferents from konio layers are thought to project¹⁴. Perhaps the upper layer 2/3 inputs are numerically weaker or have fewer terminal branches than those to lower layer 2/3 and 4A. Given that not one of 77 LGN afferents recorded in superficial cortex was red/green, however, it seems reasonable that if such a projection exists, it is either very sparse or consists of terminal arbors whose morphologies preclude successful recording.

We have shown that LGN afferents terminating in lower 2/3 and 4A are functionally distinct from those projecting to layer 4C, consistent with anatomical studies suggesting that afferents terminating in either location do not terminate in both¹⁰. Interpreted within the context of parallel anatomical projections from magno, parvo and konio LGN to V1, it seems that red/green, blue/yellow and achromatic inputs are carried by different systems, and are partitioned with these pathways to different input layers of cortex (Fig. 3b). Our observations strongly suggest that koniocellular neurons, the small cells retrogradely labelled from superficial cortex and scattered in and around konio layers, are the carriers of blue-ON and perhaps blue-OFF input to superficial layers in the macaque monkey. We found no afferents in superficial layers relaying red/green input, a role that seems to belong to the parvocellular projection to 4C β . This bears on a recent hypothesis²² that states that the parvo system is used for spatial processing only (as it is the pathway of highest visual acuity), whereas the konio system is the dedicated pathway for colour information, carrying both red/green and blue/yellow signals to a putative colour circuit in superficial cortex. Unless a completely separate population projects to upper 2/3, the segregation of red/green afferents to 4C β makes this scenario unlikely.

It is important to note that ambiguities in the definitions and properties of LGN parvo or konio cells prevent us from definitively identifying all blue/yellow neurons as konio. Ideally, konio cells might be defined as those that project to superficial layers (4A and above) and that express the neurochemical markers calbindin and α CaM kinase. But there is evidence suggesting that at least some afferents terminating in layer 4A do not express these proteins¹⁴. Thus, if defined by their superficial projections, the blue-ON and -OFF cells that we encountered are konio. If defined by their

neurochemical signature, some of the 4A input could be parvo. The highly stereotyped response properties of blue-ON inputs to layers 4A and 2/3 suggest that these inputs arise from a single population, most probably koniocellular. Blue-OFF afferents were never found above 4A and are a separate population functionally. Retrograde tracer injections encroaching on layer 4A (but not more superficially) tend to label cells in the parvo layers of LGN that do not express α CaM kinase or calbindin, possibly implicating a separate parvo pathway to 4A as the carrier of blue-OFF signals¹⁴. Retinal anatomy shows that S cones contact OFF midget bipolars²³, which in turn contact midget ganglion cells²⁴, which may be the provenance of this putative subsystem.

A final distinction awaits further work, but it seems clear that the blue-OFF pathway is anatomically and functionally distinct from both blue-ON and red/green systems, and that blue/yellow and red/green opponent systems include separate pathways terminating in distinct laminar zones in V1. $\square$

Methods

Animal preparation and data collection

Seven juvenile macaque monkeys were used in this study (three *Macaca mulatta*, four *Macaca radiata*). Each animal was anaesthetized (sufentanil citrate) and paralysed (pancuronium bromide), and dexamethasone was given to reduce brain swelling. Eyes were dilated, protected with gas-permeable contact lenses, and refracted with external lenses. We made a large ($\sim 8 \times 5$ mm) craniotomy over the parafoveal representation in striate cortex, reflected dura, and placed a $\sim 7 \times 5$ mm section of Gelfoam onto the pial surface, leaving a small gap between skull and Gelfoam to tangentially introduce an electrode. The craniotomy was sealed with bone wax and warm agar, with a tube running through to the Gelfoam to administer muscimol (50 mM in 0.9% saline, 0.1–0.2 ml h⁻¹). Muscimol perfusion inactivated all V1 neurons under the Gelfoam patch, through layer 6.

Geniculate afferent spikes were recorded with sharp, low-impedance electrodes (1–2 M Ω at 1 kHz) and sorted with custom software (PEP, D. Ringach). After encountering the first afferent 'hash' (usually lower layer 2/3 or 4A), we advanced 20–40 μ m between recordings, isolating the largest spikes present. These were most probably summed across terminal branches of single axons and not fibres of passage, because layers 4B and 5 (through which axons to superficial layers must pass) generally had no isolatable spikes and very little background activity. The large patch of inactivated cortex precludes our recording from axons of intrinsic horizontal connections. Slight shifts in receptive field position between consecutive recording sites, as well as the known convergence of many geniculate axons to any given point in the input layers of V1²⁵, argue against our consistently recording from the same axon at different sites along a penetration. See ref. 26 for a more complete discussion of the muscimol method of afferent recording. All procedures were approved by the Salk Institute Animal Care and Use Committee.

Histology

Small electrolytic lesions (3.5 μ A for 3.5 s, electrode tip negative) were made in cortex to guide the reconstruction of electrode tracks. After transcardial perfusion, the brain was blocked, sunk in 30% sucrose, and sectioned (50 μ m) on a freezing microtome. Tissue was stained for cytochrome oxidase to define layers. Owing to the length of the experiment (4–5 days) and the elimination of cortical activity, blobs could not be reliably recovered.

Visual stimuli and display calibration

Visual stimuli were generated on a Silicon Graphics O₂ computer using custom software (PEP, D. Ringach) and shown on a calibrated CRT display run at 100 Hz refresh rate. The three monitor guns were linearized with respect to frame buffer value, and their additivity verified. The monitor was placed 100 cm from the animal.

Cone-isolating stimuli were constructed from human cone fundamentals²⁷ and the measured spectral power distributions of monitor phosphors²⁸. The peak and trough of each stimulus modulated about a fixed grey background (each linearized gun at half-maximal intensity, mean luminance of ~ 28 cd m⁻², CIE colour coordinates $x = 0.29$, $y = 0.27$). Each stimulus (L, M, S, L+M and L+M+S) was presented at the maximum cone contrast achievable by our monitor, which for L (M, S) was 0.23 (0.28, 0.90). The L+M stimulus consisted of 0.55 cone contrast for L summed with 0.64 for M.

Circular patches of drifting sinusoidal gratings (radius ~ 1 – 2°) were shown at the cell's optimal temporal and spatial frequencies. Stimuli isolating different cone directions were presented in random order for 4 s each, with a blank trial inserted after every third grating. We typically averaged over 3–4 such trials per cell. In reverse-correlation experiments, Hartley stimuli²⁰ were randomly drawn and presented at every other screen refresh (every 20 ms), for a duration of ~ 15 – 20 min per cone direction.

Our calibration methods were checked by constructing cone-isolating stimuli for three sets of glass filters (L, Oriel 59500+Schott BG-40; M, Schott VG-9; S, Oriel 59814+Oriel 59080) whose transmissions approximated the spectral absorptions of L, M and S cones²⁹. We measured the power of these cone-isolating stimuli through each filter set with a photodiode, and found that each stimulus isolated its intended model cone at least 30 times better than the other two model cones. In addition, the multiple physiological controls in a previous study by the authors³⁰ (including controls for L/M/rod intrusion

into S-cone responses using a very bright yellow adapting light, as well as tests for chromatic aberration by varying the spatial frequency of cone-isolating stimuli from near full-field to optimal) apply to the present study, because apparatus, set-up and calibration methods were identical.

Received 16 June; accepted 15 October 2003; doi:10.1038/nature02163.

1. Lennie, P. & D'Zmura, M. Mechanisms of color vision. *Crit. Rev. Neurobiol.* **3**, 333–400 (1988).
2. Dacey, D. M. Parallel pathways for spectral coding in primate retina. *Annu. Rev. Neurosci.* **23**, 743–775 (2000).
3. De Valois, R. L., Cottaris, N. P., Elfar, S. D., Mahon, L. E. & Wilson, J. A. Some transformations of color information from lateral geniculate nucleus to striate cortex. *Proc. Natl Acad. Sci. USA* **97**, 4997–5002 (2000).
4. Dacey, D. M., Peterson, B. B., Robinson, F. R. & Gamlin, P. D. Fireworks in the primate retina. *In vitro* photodynamics reveals diverse LGN-projecting ganglion cell types. *Neuron* **37**, 15–27 (2003).
5. Leventhal, A. G., Rodieck, R. W. & Dreher, B. Retinal ganglion cell classes in the Old World monkey: Morphology and central projections. *Science* **213**, 1139–1142 (1981).
6. Michael, C. R. Retinal afferent arborization patterns, dendritic field orientations, and the segregation of function in the lateral geniculate nucleus of the monkey. *Proc. Natl Acad. Sci. USA* **85**, 4914–4918 (1988).
7. Conley, M. & Fitzpatrick, D. Morphology of retinogeniculate axons in the macaque. *Vis. Neurosci.* **2**, 287–296 (1989).
8. Hubel, D. & Wiesel, T. Laminar and columnar distribution of geniculocortical fibers in the macaque monkey. *J. Comp. Neurol.* **146**, 421–450 (1972).
9. Hendrickson, A. E., Wilson, J. R. & Ogren, M. P. The neuroanatomical organization of pathways between the dorsal lateral geniculate nucleus and visual cortex in Old World and New World primates. *J. Comp. Neurol.* **182**, 123–136 (1978).
10. Blasdel, G. G. & Lund, J. S. Termination of afferent axons in macaque striate cortex. *J. Neurosci.* **3**, 1389–1413 (1983).
11. Rodieck, R. W. & Watanabe, M. Survey of the morphology of macaque retinal ganglion cells that project to the pretectum, superior colliculus, and parvocellular laminae of the lateral geniculate nucleus. *J. Comp. Neurol.* **338**, 289–303 (1993).
12. Martin, P. R., White, A. J., Goodchild, A. K., Wilder, H. D. & Sefton, A. E. Evidence that blue-on cells are part of the third geniculocortical pathway in primates. *Eur. J. Neurosci.* **9**, 1536–1541 (1997).
13. Hendry, S. H. & Reid, R. C. The koniocellular pathway in primate vision. *Annu. Rev. Neurosci.* **23**, 127–153 (2000).
14. Hendry, S. H. & Yoshioka, T. A neurochemically distinct third channel in the macaque dorsal lateral geniculate nucleus. *Science* **264**, 575–577 (1994).
15. Wiesel, T. N. & Hubel, D. H. Spatial and chromatic interactions in the lateral geniculate body of the rhesus monkey. *J. Neurophysiol.* **29**, 1115–1156 (1966).
16. Schiller, P. H. & Malpel, J. G. Functional specificity of lateral geniculate nucleus laminae of the rhesus monkey. *J. Neurophysiol.* **41**, 788–797 (1978).
17. Derrington, A. M., Krauskopf, J. & Lennie, P. Chromatic mechanisms in lateral geniculate nucleus of macaque. *J. Physiol. (Lond.)* **357**, 241–265 (1984).
18. Reid, R. C. & Shapley, R. M. Space and time maps of cone photoreceptor signals in macaque lateral geniculate nucleus. *J. Neurosci.* **22**, 6158–6175 (2002).
19. Valberg, A., Lee, B. B. & Tigwell, D. A. Neurons with strong inhibitory S-cone inputs in the macaque lateral geniculate nucleus. *Vision Res.* **26**, 1061–1064 (1986).
20. Ringach, D. L., Sapiro, G. & Shapley, R. A subspace reverse-correlation technique for the study of visual neurons. *Vision Res.* **37**, 2455–2464 (1997).
21. Dacey, D. M. & Lee, B. B. The 'blue-on' opponent pathway in primate retina originates from a distinct bistratified ganglion cell type. *Nature* **367**, 731–735 (1994).
22. Calkins, D. J. & Sterling, P. Evidence that circuits for spatial and color vision segregate at the first retinal synapse. *Neuron* **24**, 313–321 (1999).
23. Ahmad, K. M., Klug, K., Herr, S., Sterling, P. & Schein, S. Cell density ratios in a foveal patch in macaque retina. *Vis. Neurosci.* **20**, 189–209 (2003).
24. Klug, K., Herr, S., Ngo, I. T., Sterling, P. & Schein, S. J. Macaque retina contains an S-cone OFF midgate pathway. *J. Neurosci.* **23**, 9881–9887 (2003).
25. Blasdel, G. G. & Fitzpatrick, D. Physiological organization of layer 4 in macaque striate cortex. *J. Neurosci.* **4**, 880–895 (1984).
26. Chapman, B., Zahs, K. R. & Stryker, M. P. Relation of cortical cell orientation selectivity to alignment of receptive fields of the geniculocortical afferents that arborize within a single orientation column in ferret visual cortex. *J. Neurosci.* **11**, 1347–1358 (1991).
27. Stockman, A., MacLeod, D. I. & Johnson, N. E. Spectral sensitivities of the human cones. *J. Opt. Soc. Am. A* **10**, 2491–2521 (1993).
28. Wandell, B. A. *Foundations of Vision* 413–421 (Sinauer, Sunderland, MA, 1995).
29. Chichilnisky, E. J. & Baylor, D. A. Receptive-field microstructure of blue-yellow ganglion cells in primate retina. *Nature Neurosci.* **2**, 889–893 (1999).
30. Chatterjee, S. & Callaway, E. M. S cone contributions to the magnocellular visual pathway in macaque monkey. *Neuron* **35**, 1135–1146 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank D. Ringach for providing software used for visual stimulation, spike sorting, and some data analysis; E. J. Chichilnisky for help with stimulus calibration and design; and E. J. Chichilnisky and G. Horwitz for a critical reading of the manuscript. We also thank S. Tye for surgical assistance.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.C. (sochatte@ucsd.edu).

Development and maintenance of B and T lymphocytes requires antiapoptotic MCL-1

Joseph T. Opferman, Anthony Letai, Caroline Beard, Mia D. Sorcinelli, Christy C. Ong & Stanley J. Korsmeyer

Howard Hughes Medical Institute, Dana Farber Cancer Institute, Department of Pathology and Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA

Regulated apoptosis is essential for both the development and the subsequent maintenance of the immune system^{1,2}. Interleukins, including IL-2, IL-4, IL-7 and IL-15, heavily influence lymphocyte survival during the vulnerable stages of VDJ rearrangement and later in ensuring cellular homeostasis, but the genes specifically responsible for the development and maintenance of lymphocytes have not been identified^{3–8}. The antiapoptotic protein MCL-1 is an attractive candidate, as it is highly regulated⁹, appears to enhance short-term survival¹⁰ and functions at an apical step in genotoxic deaths¹¹. However, *Mcl-1* deficiency results in peri-implantation lethality¹². Here we show that mice conditional for *Mcl-1* display a profound reduction in B and T lymphocytes when MCL-1 is removed. Deletion of *Mcl-1* during early lymphocyte differentiation increased apoptosis and arrested the development at pro-B-cell and double-negative T-cell stages. Induced deletion of *Mcl-1* in peripheral B- and T-cell populations resulted in their rapid loss. Moreover, IL-7 both induced and required MCL-1 to mediate lymphocyte survival. Thus, MCL-1, which selectively inhibits the proapoptotic protein BIM, is essential both early in lymphoid development and later on in the maintenance of mature lymphocytes.

The proapoptotic 'BH3-only' members of the BCL-2 family respond to selective death signals and trigger activation of the 'multidomain' death effectors BAX and BAK, which constitute an obligate gateway to the intrinsic death pathway. Conversely, antiapoptotic BCL-2 members have an important role in binding and sequestering BH3-only molecules, thus preventing activation of BAX and BAK^{13,14}. BH3 domains can be subdivided into those that 'activate' (for example, BID and BIM) by inducing oligomerization of BAX and BAK, or those that 'sensitize' (for example, BAD and BIK) by occupying the pocket of antiapoptotic family members¹⁵. Individual BCL-2 family members have specific roles. Haematopoiesis is initially normal in *Bcl-2*-deficient animals, but over time lymphocytes are vulnerable to apoptosis, particularly following activation stimuli¹⁶. *Bcl-X_L*-deficient lymphocytes show that it is critical for the survival of CD4⁺CD8⁺ double-positive (DP) thymocytes, but not for mature T cells^{17,18}.

We generated a conditional *Mcl-1* allele by targeting *loxP* sites upstream of the ATG start codon and between exons 1 and 2 (Fig. 1a). This *Mcl-1*^{lox(f)} allele was transmitted through the germ line, and matings of *Mcl-1*^{f/wt} mice yielded viable *Mcl-1*^{f/f} offspring at the expected mendelian frequency (Supplementary Fig. 1a–d). Both *Mcl-1*^{f/wt} and *Mcl-1*^{f/f} cells express levels of MCL-1 protein comparable to wild-type (wt) cells (Fig. 1b and Supplementary Fig. 3a). We also generated an *Mcl-1*^{null} allele in which portions of *Mcl-1* exons 1 and 2 were replaced with an internal ribosome entry site allowing the expression of a neomycin-β-galactosidase fusion protein (Fig. 1a). *Mcl-1*^{f/null} murine embryonic fibroblasts (MEFs) after Cre-mediated deletion, *Mcl-1*^{deleted/null} (Δ /null), lack detectable MCL-1 protein (Fig. 1b).

To restrict deletion of the *Mcl-1*^f allele to T-cell development, we introduced the lymphocyte-specific protein tyrosine kinase–Cre

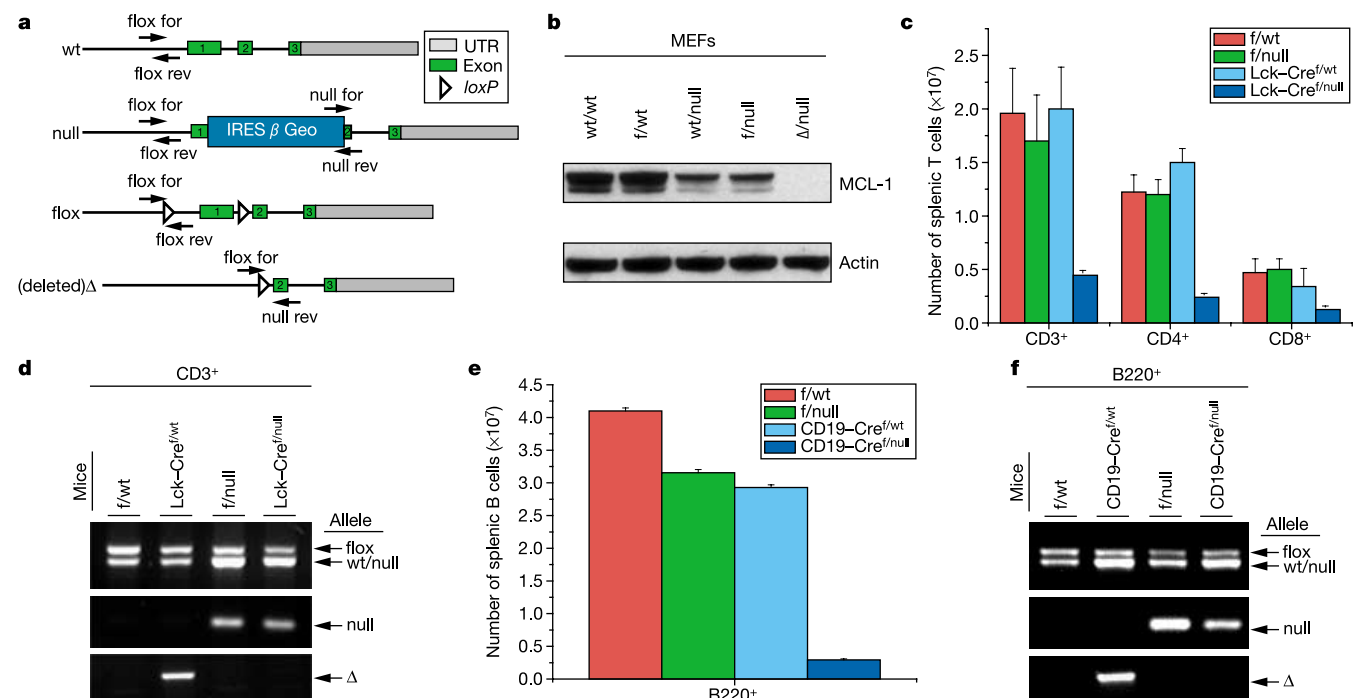


Figure 1 Conditional deletion of *Mcl-1* results in loss of peripheral lymphocytes.

a, Schematic of *Mcl-1* loci. PCR primers distinguishing these alleles include flox-forward (flox for) and flox-reverse (flox rev) primers that generate a 350-base-pair (bp) product in *wt* and *null* alleles, but a 400-bp product in *flox* alleles owing to the *loxP* site. The null primer set amplifies only the *Mcl-1*^{null} allele, whereas a combination of flox-for and null-rev primers amplifies selectively the *Mcl-1*^{deleted} allele. UTR, untranslated region.

b, Immunoblots of lysates from murine embryonic fibroblasts (MEFs). *Mcl-1*^Δ cell lines were generated by transduction of tyrosine aminotransferase (Tat)-Cre fusion protein.

c, Number of splenic T-cell subsets (mean ± s.e.m.) for three mice of *Mcl-1*^{f/wt}, *Mcl-1*^{f/null}, *Lck-CreMcl-1*^{f/wt} or *Lck-CreMcl-1*^{f/null} genotype at 8 weeks of age.

d, Genomic DNA from sorted CD3⁺ splenic T cells was analysed by PCR to determine the representation of *Mcl-1*^{f/wt}, *Mcl-1*^{flox}, *Mcl-1*^Δ and *Mcl-1*^{null} alleles.

e, Number of splenic B220⁺ cells (mean ± s.e.m.) for three mice of *Mcl-1*^{f/wt} and *Mcl-1*^{f/null} with and without CD19-Cre^{f/wt} and CD19-Cre^{f/null} genotypes at 8 weeks of age.

f, Genomic DNA from sorted B220⁺ splenic B cells was analysed by PCR to determine the representation of *Mcl-1* alleles.

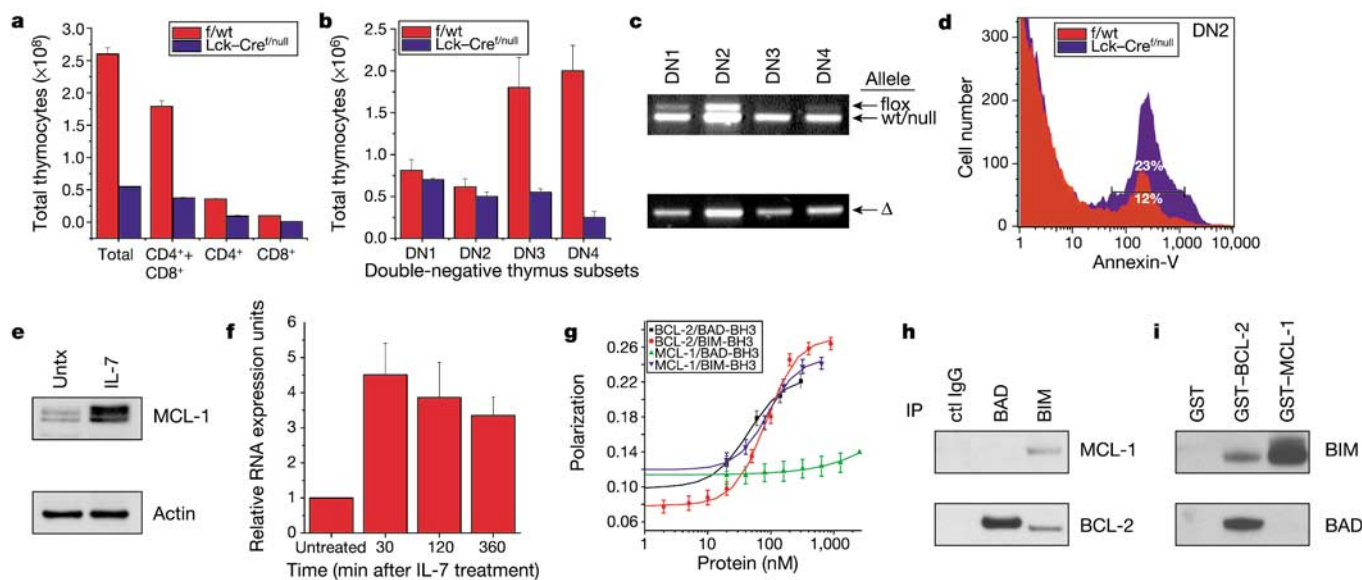


Figure 2 Lck-Cre-mediated deletion of *Mcl-1* blocks thymocyte development. **a**, Number of total thymocytes (mean ± s.e.m.) from three mice at 8 weeks of age of *f/wt* or *Lck-CreMcl-1*^{f/null}. **b**, Total DN thymocyte subsets (mean ± s.e.m.) from three mice of the respective genotypes at 8 weeks of age. **c**, Genomic DNA from sorted DN thymocyte subsets of *Lck-CreMcl-1*^{f/wt} mice was analysed by PCR to identify the onset of the *Mcl-1*^Δ allele during development. **d**, Percentage apoptosis of the DN2 thymocytes was assessed with Annexin-V for the *f/wt* and *Lck-CreMcl-1*^{f/null} genotypes. Histograms are representative of three mice analysed at 8 weeks of age. **e**, MCL-1 protein levels as determined by immunoblots of cultured thymocytes from *Rag-2*-deficient mice in the absence (Untx) or presence of

IL-7 (10 ng ml⁻¹) for 16 h. **f**, *Mcl-1* mRNA levels in cultured *Rag-2*-deficient thymocytes after 30, 120 and 360 min of exposure to 10 ng ml⁻¹ IL-7. The mean ± s.e.m. is presented of triplicate assays in which *Mcl-1* mRNA expression units were normalized to *HPRT*. **g**, Isotherms of fluorescein-tagged BIM-BH3 or BAD-BH3 peptides binding to GST-BCL-2 or GST-MCL-1 quantified by fluorescence polarization. Error bars denote s.d. from the mean. **h**, Mouse thymocyte lysates were immunoprecipitated (IP) with anti-BAD or anti-BIM antibodies; immunoprecipitates were resolved and immunoblots developed with indicated antibodies. **i**, Mouse thymocyte lysates were incubated with GST-BCL-2 or GST-MCL-1 fusion proteins, complexes resolved and immunoblots developed as indicated.

(Lck-Cre) transgene¹⁹. Lck-Cre*Mcl-1*^{wt/null} mice were crossed with *Mcl-1*^{fl/fl} mice and the progeny analysed. Peripheral T lymphocytes were decreased in both spleen and lymph nodes of Lck-Cre*Mcl-1*^{fl/null} mice when compared with littermate control mice (*f/wt*, *f/null* and Lck-Cre*Mcl-1*^{fl/wt} all showed normal numbers) (Fig. 1c and Supplementary Fig. 2a, c). Both CD4⁺ and CD8⁺ subsets were depleted in these Lck-Cre*Mcl-1*^{fl/null} mice (Fig. 1c and Supplementary Fig. 2b). To determine whether the remaining T cells survived without MCL-1 or represent a minority population that failed to delete, we used polymerase chain reaction (PCR) assays that distinguish *Mcl-1*^f, *Mcl-1*^Δ and *Mcl-1*^{null} alleles (Fig. 1a, d). The deleted allele, *Mcl-1*^Δ, was detectable only in CD3⁺ T cells bearing a *Mcl-1*^{wt} allele and was absent in the residual T cells of Lck-Cre*Mcl-1*^{fl/null} mice (Fig. 1d). Consistent with this observation, the intact *Mcl-1*^f allele was present in this residual T-cell population. This suggests that T cells lacking both copies of *Mcl-1* are non-viable, and that the residual T cells found in Lck-Cre*Mcl-1*^{fl/null} mice exist because they have escaped Cre-mediated deletion.

CD19-Cre*Mcl-1*^{fl/null} mice with Cre recombinase under the control of a B-lineage-restricted gene²⁰ exhibited a profound reduction in peripheral B cells when compared with littermates with intact *Mcl-1* (Fig. 1e and Supplementary Fig. 2d, e). *Mcl-1* deletion reduced the number of splenic B cells to 12% of the levels in *Mcl-1*^{fl/wt} mice, whereas lymph node B cells were barely detectable. Hemizygosity for *Mcl-1* (*Mcl-1*^{fl/null}, CD19-Cre*Mcl-1*^{fl/wt}) resulted in reduced peripheral B220⁺ cells, indicating that *Mcl-1* dosage has more effect on B-cell lineages than it does on T-cell lineages (Fig. 1e

and Supplementary Fig. 2e). The few remaining peripheral B cells in CD19-Cre*Mcl-1*^{fl/null} mice escaped deletion as they displayed detectable *Mcl-1*^f alleles but no *Mcl-1*^Δ alleles by PCR (Fig. 1f). This suggests that MCL-1 is necessary for the development or maintenance of mature B cells.

We next examined whether MCL-1 was required at earlier stages of lymphocyte development. The total number of thymocytes was markedly reduced in Lck-Cre*Mcl-1*^{fl/null} mice (Fig. 2a; <20% of littermate *Mcl-1*^{fl/wt} mice). The numbers of CD4⁺CD8⁺ DP cells, as well as CD4⁺ or CD8⁺ single-positive (SP) cells, were all decreased (Fig. 2a and Supplementary Fig. 3b). This warranted an examination of earlier CD4⁺CD8⁺ double-negative (DN) stages of thymocyte development that can be subdivided into four serial subclasses. Developing pro-T cells are initially CD44⁺CD25[−] DN stage 1 (DN1), progress to CD44⁺CD25⁺ DN stage 2 (DN2), followed by CD44⁺CD25⁺ DN stage 3 (DN3) and finally CD44⁺CD25⁺ DN stage 4 (DN4) before becoming DP thymocytes²¹. The total number of DN1 and DN2 thymocytes was relatively normal in Lck-Cre*Mcl-1*^{fl/null} mice (Fig. 2b). By contrast, the DN3 and DN4 subsets were markedly reduced on deletion of *Mcl-1* (Fig. 2b and Supplementary Fig. 3c). The deleted *Mcl-1*^Δ allele was present in all four DN stages of development (Fig. 2c), consistent with studies placing the initiation of Lck-Cre-mediated deletion at the onset of CD44 expression¹⁹. Because DN3 and DN4 thymocytes are markedly depleted on deletion of *Mcl-1*, we examined the cell fate of the previous DN2 thymocytes and found a significant increase in the percentage of apoptotic (Annexin-V⁺) DN2 cells (Fig. 2d).

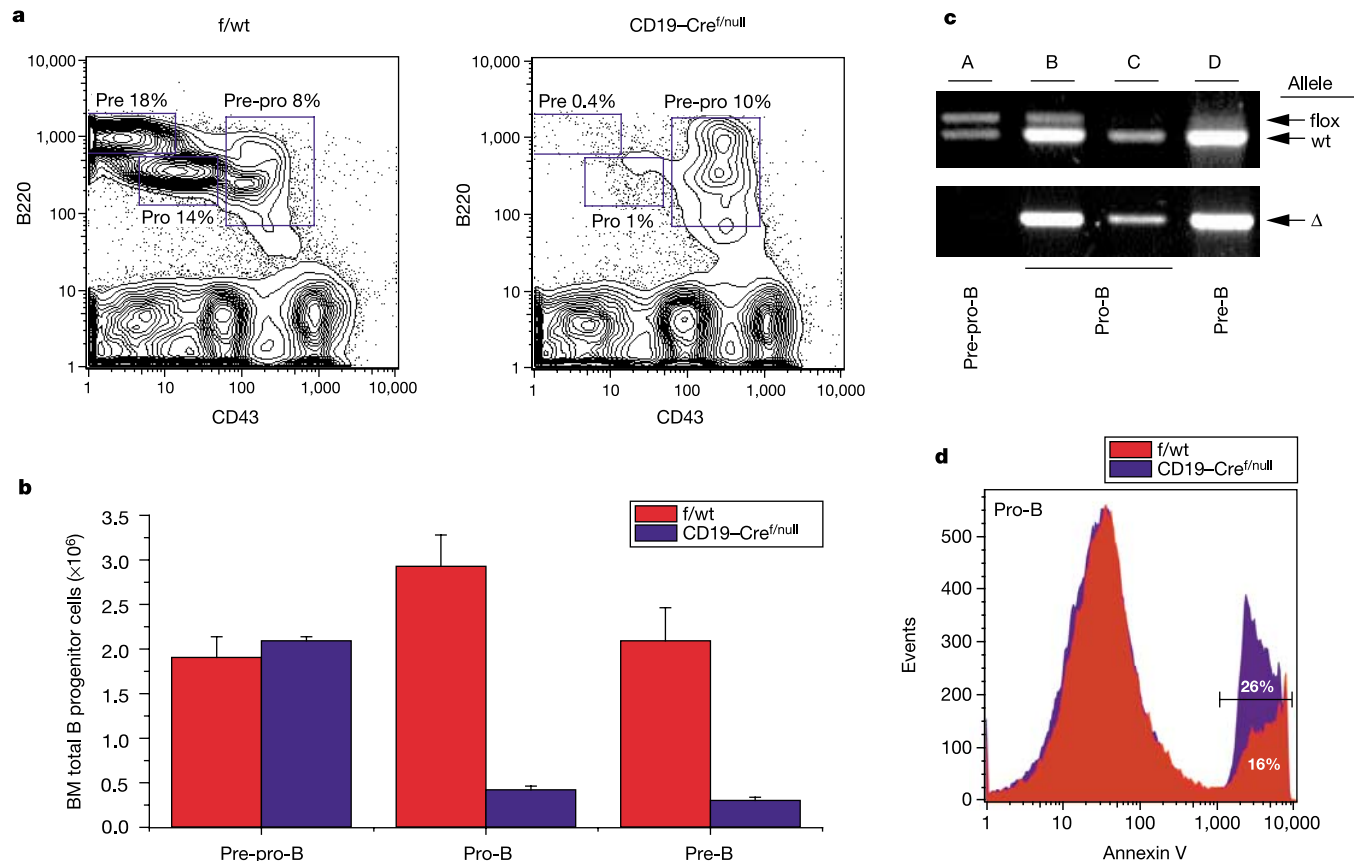


Figure 3 CD19-Cre-mediated deletion of *Mcl-1* blocks B-cell development. **a**, The percentage of bone marrow pre-pro-B (CD43⁺B220⁺, right gate), pro-B (CD43⁺B220⁺, middle gate) and pre-B (CD43⁺B220⁺, left gate) subsets as determined by flow cytometry for *f/wt* and CD19-Cre^{f/null} genotypes. **b**, Total B-cell progenitors (mean ± s.e.m.) recovered from the bone marrow of *f/wt* and CD19-Cre^{f/null} genotypes

at 8 weeks of age. **c**, Genomic DNA from sorted cells representing Hardy subsets (A, B, C, D) of B-cell progenitors from CD19-Cre^{f/wt} mice was analysed by PCR to detect the *Mcl-1*^Δ allele. **d**, Percentage apoptosis of pro-B (CD43⁺B220⁺) cells from *f/wt* or CD19-Cre^{f/null} mice was assessed with Annexin-V. Histograms are representative of three mice analysed at 8 weeks of age.

Although deletion of *Mcl-1* starts by DN1, the most prominent block manifests at DN2/DN3. It should be noted that DN2 cells undergo T-cell receptor rearrangement and are highly dependent on cytokine signalling for survival²².

The three main stages of early B-cell development are: pre-pro-B (CD43⁺B220⁺CD19⁻; subfraction A), pro-B (CD43⁺B220⁺CD19⁺; subfractions B and C) and pre-B (CD43⁻B220⁺CD19⁺; subfraction D)²³. The percentage and absolute number of pre-pro-B cells in CD19-Cre*Mcl-1*^{f/null} bone marrow were comparable to those in littermates that retained *Mcl-1* (Fig. 3a, b). However, the pro-B- and pre-B-cell populations were markedly reduced after *Mcl-1* deletion (Fig. 3b and Supplementary Fig. 4). This includes the earliest of the pro-B-cell stages, subset B, which was depleted, as were all subsequent stages (Fig. 3a, b and Supplementary Fig. 4). Although the pro-B, subfraction A cells had not deleted *Mcl-1*, stages B, C and D all showed *Mcl-1*^Δ (Fig. 3c). Subfraction B pro-B cells with *Mcl-1*^Δ showed more apoptotic (Annexin-V⁺) cells, consistent with an increased death of pro-B cells contributing to the loss of subsequent stages (Fig. 3d).

The developmental stages at which T- and B-cell development is impaired in *Mcl-1* conditional knockouts is similar to defects observed in interleukin-7 (IL-7) or IL-7 receptor (IL-7R)-deficient

mice^{4,5}. To assess whether IL-7 can affect MCL-1 expression, we cultured recombination-activating gene 2 (*Rag-2*)-deficient thymocytes (enriched for DN1/DN2 owing to a lack of T-cell receptor rearrangement) in the presence or absence of IL-7 for 16 h (ref. 24). Treatment with IL-7 resulted in a ~6-fold increase in MCL-1 protein levels in these thymocytes (Fig. 2e). Moreover, when real-time PCR was used to quantify messenger RNA, addition of IL-7 (10 μg ml⁻¹) resulted in a ~5-fold increase in *Mcl-1* mRNA (Fig. 2f).

We addressed the question of whether mature lymphocytes also require MCL-1 for their maintenance. MCL-1 protein is present in peripheral B220⁺, CD4⁺ and CD8⁺ lymphocytes (Supplementary Fig. 5). The viability of wild-type T cells was promoted best by culture in IL-7, to a lesser extent with IL-15 and least with IL-2 (Fig. 4a). An increase in MCL-1 protein levels was most marked for T cells cultured in IL-7, followed by IL-15 and then IL-2 (Fig. 4b). Real-time PCR assay showed a quantitative increase in *Mcl-1* mRNA on exposure of mature T cells to IL-7, indicating that there is a transcriptional component to this response (Fig. 4c). To assess whether *Mcl-1* is required for IL-7-signalled survival, we turned to the MxCre recombinase model²⁵. Interferon-α (IFN-α)-induced deletion of *Mcl-1* in cultured CD3⁺ T cells from MxCre*Mcl-1*^{f/null}

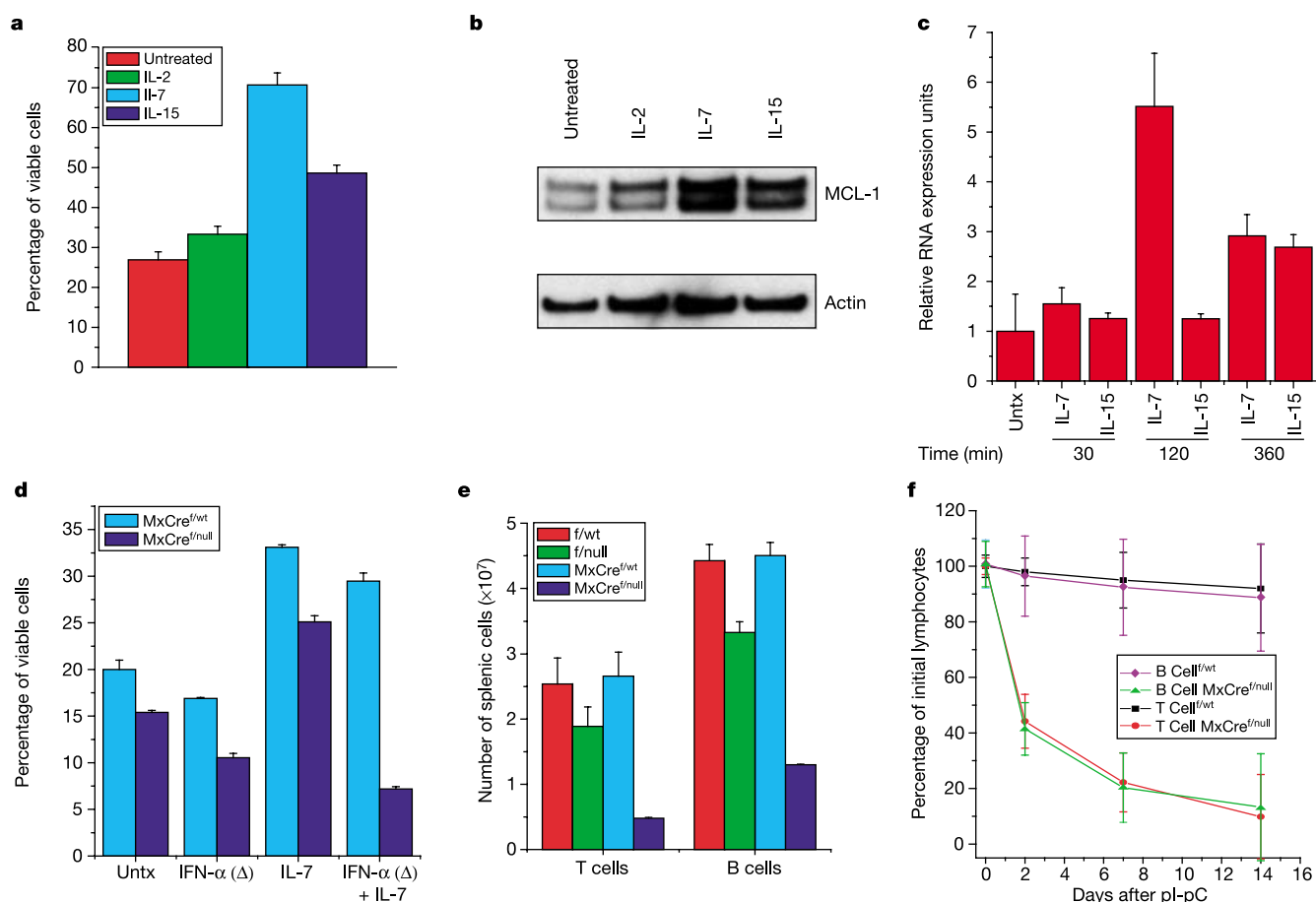


Figure 4 MCL-1 is required for mature lymphocyte survival. **a**, Viability of T cells cultured with or without cytokines (10 ng ml⁻¹) for 24 h. Mean ± s.e.m. percentage viability of triplicate experiments as determined by Annexin-V. **b**, MCL-1 protein levels in T cells after 24 h culture (10 ng ml⁻¹ cytokines) as determined by immunoblot. **c**, Mean ± s.e.m. of triplicate assays for *Mcl-1* mRNA levels normalized to *HPRT* for cultured purified T lymphocytes after 30, 120 and 360 min of exposure to cytokines (10 ng ml⁻¹). **d**, Mean ± s.e.m. percentage viability of triplicate experiments for purified splenic T cells from MxCre*Mcl-1*^{f/wt} or MxCre*Mcl-1*^{f/null} after 48 h of culture with or without IFN-α to

induce deletion and/or IL-7. **e**, Number of total splenic cells (mean ± s.e.m.) from three mice of each genotype at 6 weeks of age, assessed 7 days after pl-pC treatment. **f**, *Mcl-1*^{f/null} or MxCre*Mcl-1*^{f/null} purified T or B lymphocytes were adoptively transferred to *Rag-2*-deficient recipients. After engraftment, pl-pC was delivered to delete the *Mcl-1*^{fllox} allele. Mice were analysed at day 0 (before pl-pC), day 2 and day 7 to determine CD3⁺ and B220⁺ lymphocytes in the spleen and lymph nodes. The mean ± s.e.m. represents the percentage of initial B or T cells recovered in recipient mice at day 0, for three recipient mice at each time point.

mice revealed that the IL-7-mediated survival of mature T cells requires MCL-1 (Fig. 4d).

To address whether MCL-1 is required for the maintenance of existing, mature lymphocytes *in vivo*, we administered polyinosinic-polycytidylic acid (pI-pC) to MxCreMcl-1^{fnull} mice, which induces deletion of *Mcl-1*. Both B and T cells were depleted from the spleen 7 days after administration of pI-pC (Fig. 4e). As MxCre deletes *Mcl-1* in multiple cell types, we performed an adoptive transfer to assess whether the loss of mature B and T cells represented a cell-autonomous event. CD3⁺ and B220⁺ peripheral lymphocytes were enriched by bead depletion of myeloid and natural killer cells from the spleen and lymph nodes of MxCreMcl-1^{fnull} mice. Mature lymphocytes (5 × 10⁶) were adoptively transferred to Rag-2-deficient recipients and allowed to engraft over 7–10 days. pI-pC was then administered to induce deletion of *Mcl-1* in the donor lymphocytes. Within 2 days of pI-pC injection, the percentage of both T and B lymphocytes in spleen and lymph nodes was reduced by ~60% for MxCreMcl-1^{fnull} cells compared with Mcl-1^{fwt} cells (Fig. 4f). Peripheral B and T cells were progressively lost over the 2-week time course following deletion of *Mcl-1* (Fig. 4f). This adoptive transfer regimen indicates an intrinsic requirement for MCL-1 in the maintenance of peripheral B and T cells, which are rapidly lost on *Mcl-1* deletion.

To examine why MCL-1 is singularly required at cytokine-dependent steps, we assessed its capacity to bind selected BH3-only members. Two BH3-only members, BIM and BAD, are involved in cytokine-dependent survival pathways. *Bim*-deficient cells display resistance to death due to cytokine deprivation²⁶; by contrast, *Bad*-null mice and knock-in mice bearing a non-phosphorylatable BAD^{35A} molecule show an altered apoptotic threshold for cytokine deprivation, confirming a 'sensitizer' role for BAD^{27,28}. Recombinant MCL-1 protein showed no substantial binding to the BAD-BH3 peptide (>2,900 nM), but did display a strong interaction with BIM-BH3 (74 ± 2 nM) (Fig. 2g). By contrast, BCL-2 showed comparable binding to both BAD and BIM-BH3 peptides¹⁵ (Fig. 2g). Consistent with these determined affinities, immunoprecipitation of proteins from solubilized thymocytes revealed that MCL-1 interacts with BIM not BAD (Fig. 2h). This was confirmed with glutathione S-transferase (GST)–MCL-1 protein, which pulled down substantially more BIM from thymocyte extracts than did comparable GST–BCL2; however, MCL-1 again did not associate with BAD (Fig. 2i). This suggests that there are meaningful subsets of antiapoptotic BCL-2 members based on their selective interactions, just as there are biologically relevant subsets of proapoptotic members. A rationale for the importance of MCL-1 would be its selective sequestration of the 'activator' BH3-only BIM, thus preventing the activation of BAX and BAK.

Unexpectedly, MCL-1 showed dual requirements in the well-defined differentiation pathways of lymphocytes. Mature T and B cells have a striking dependence on MCL-1 for their survival. In addition, loss of MCL-1 increased apoptosis at DN2 and arrested the development of DN3 thymocytes, as well as arresting development of comparable-stage pro-B cells (subfraction B). Note that both DN2 and pro-B cells depend on the cytokine IL-7 signalling through common γ -chain and IL-7R α chain pathways^{3–5}. Overexpression of antiapoptotic BCL-2 has been shown to rescue T-cell development in IL-7R knockout mice^{29,30}, indicating that one of the main effects of IL-7 is to promote cell survival. IL-7 substantially increased the levels of MCL-1 in immature thymocytes and mature T cells, and enhanced their survival. Early lymphocyte development and mature lymphocyte homeostasis share the common characteristic of a cytokine-dependent step that is totally reliant on MCL-1 to block apoptosis. The evidence supports a model in which MCL-1 is required at a distinct apical step¹¹ to counter BIM selectively, a checkpoint that could prove essential for multiple death/survival pathways. □

Methods

Targeting the *Mcl-1* genomic locus

The *Mcl-1* genomic locus was targeted by placing a *loxP* site upstream of the ATG start codon. A *loxP*-flanked cytomegalovirus (CMV) promoter driving the expression of a hygromycin–thymidine kinase fusion protein was inserted in the intron between exons 1 and 2. RW4 embryonic stem cells (129SvJ) were targeted with the construct, followed by transient transfection with CMV–Cre. Embryonic stem cell clones were selected and screened for successful recombination, and those containing two *loxP* sites were injected into C57BL/6 blastocysts.

Cell sorting and viability

Thymus, spleen, lymph nodes and bone marrow suspensions were prepared from animals that were killed at 6–12 weeks of age. Stained cells were analysed on a FACSCalibur (Becton Dickinson). For high-speed cell sorting, cells were analysed and collected on a MoFlow cell sorter (MoFlow). Cell-surface stains were purchased from BD-Pharmingen, with the exception of anti-CD127 (Ebioscience). Bead enrichment (>90%) for desired peripheral lymphocytes was performed by positive selection using magnetic beads (Dyna and Miltenyi). Viability was assessed by flow cytometric analysis of Annexin-V and propidium iodide staining (BioVision).

Analysis of MCL-1 protein

Rag-2-deficient thymocytes (6–8 weeks of age) were cultured with or without recombinant mouse IL-7 (R&D). After 16 h of culture, cells were lysed and proteins were separated by 10% SDS–PAGE (Invitrogen); they were then transferred and immunoblotted using an MCL-1-specific rabbit polyclonal generated against MCL-1 peptide (sequence, SPEELDGCEPAIGKRPV) and anti- β -actin (Chemicon).

Analysis of cytokine treatment on *Mcl-1* mRNA expression

Rag-2-deficient thymocytes or purified T lymphocytes were starved in complete RPMI (Roswell Park Memorial Institute) media with 10% fetal calf serum for 2 h. IL-7 was then re-administered to the cultures and cells were harvested at 30, 120 and 160 min. Total RNA was extracted (Trizol) and complementary DNA synthesized (Promega) to act as template for TaqMan real-time PCR analysis (Applied Biosystems, Prism). PCR primers were *Mcl-1* cDNA (forward, AGAGCGCTGGAGACCCTG; reverse, CTATCTTATTAGATATGCCAGACC) and hypoxanthine phosphoribosyl transferase (*HPRT*) (forward, GTTGGATACAGGCCAGACTTTGTTG; reverse, GAGGGTAGGCTGGCTATAGGCT). These primer sets were used with Sybr-Green (Applied Biosystems) to measure *Mcl-1* relative transcript levels and *HPRT* to control for input cDNA.

In vitro MxCre deletion

CD3⁺ peripheral lymphocytes were enriched from MxCre^{fwt} plus MxCre^{fnull} mice. The cells were cultured with or without IFN- α (PBL Biomedical) or IL-7, or with both. After 48 h, the cells were analysed by flow cytometry.

Peripheral *Mcl-1* deletion and analysis

Lymphocytes were enriched by bead depletion of myeloid and natural killer cells from the spleen and lymph nodes of MxCreMcl-1^{fwt} and MxCreMcl-1^{fnull} mice. Lymphocytes (5 × 10⁶) were adoptively transferred to Rag-2-deficient recipients (6–12 weeks of age). After 7–10 days of engraftment, mice were injected with pI-pC (400 μ g) to stimulate IFN- α , inducing Cre. Recipient mice were killed and analysed.

BH3-only interaction experiments

Affinities of interaction between GST fusion proteins and BH3 peptides were determined as previously described by fluorescence polarization¹⁵. Co-immunoprecipitations were performed on whole-cell, wild-type thymocytes lysed in 0.2% NP-40 using anti-BAD (clone 48, BD-Transduction Labs) and anti-BIM (clone 14A8, Chemicon), then immunoblotted as indicated. GST pull-down experiments were performed on thymocyte lysates (0.2% NP-40) then immunoblotted.

Received 28 August; accepted 22 September 2003; doi:10.1038/nature02067.

- Marsden, V. S. & Strasser, A. Control point of apoptosis in the immune system: Bcl-2, BH3-only proteins and more. *Annu. Rev. Immunol.* **21**, 71–105 (2003).
- Rathmell, J. C. & Thompson, C. B. The central effectors of cell death in the immune system. *Annu. Rev. Immunol.* **17**, 781–828 (1999).
- Cao, X. *et al.* Defective lymphoid development in mice lacking expression of the common cytokine receptor γ chain. *Immunity* **2**, 223–238 (1995).
- Peschon, J. J. *et al.* Early lymphocyte expansion is severely impaired in interleukin 7 receptor-deficient mice. *J. Exp. Med.* **180**, 1955–1960 (1994).
- von Freeden-Jeffry, U. *et al.* Lymphopenia in interleukin (IL)-7 gene-deleted mice identifies IL-7 as a nonredundant cytokine. *J. Exp. Med.* **181**, 1519–1526 (1995).
- Vella, A. T., Dow, S., Potter, T. A., Kappler, J. & Marrack, P. Cytokine-induced survival of activated T cells *in vitro* and *in vivo*. *Proc. Natl Acad. Sci. USA* **95**, 3810–3815 (1998).
- Zhang, X., Sun, S., Hwang, I., Tough, D. F. & Sprent, J. Potent and selective stimulation of memory-phenotype CD8⁺ T cells *in vivo* by IL-15. *Immunity* **8**, 591–599 (1998).
- Schluns, K. S., Kieper, W. C., Jameson, S. C. & Lefrançois, L. Interleukin-7 mediates the homeostasis of naive and memory CD8 T cells *in vivo*. *Nature Immunol.* **1**, 426–432 (2000).
- Yang, T., Buchan, H. L., Townsend, K. J. & Craig, R. W. MCL-1, a member of the BCL-2 family, is induced rapidly in response to signals for cell differentiation or death, but not to signals for cell proliferation. *J. Cell. Physiol.* **166**, 523–536 (1996).
- Zhou, P., Qian, L., Kozopas, K. M. & Craig, R. W. Mcl-1, a Bcl-2 family member, delays the death of hematopoietic cells under a variety of apoptosis-inducing conditions. *Blood* **89**, 630–643 (1997).

11. Nijhawan, D. *et al.* Elimination of Mcl-1 is required for the initiation of apoptosis following ultraviolet irradiation. *Genes Dev.* **17**, 1475–1486 (2003).
12. Rinkenberger, J. L., Horning, S., Klocke, B., Roth, K. & Korsmeyer, S. J. Mcl-1 deficiency results in perimplantation embryonic lethality. *Genes Dev.* **14**, 23–27 (2000).
13. Wei, M. C. *et al.* Proapoptotic BAX and BAK: a requisite gateway to mitochondrial dysfunction and death. *Science* **292**, 727–730 (2001).
14. Cheng, E. H. *et al.* BCL-2, BCL-X_L sequester BH3 domain-only molecules preventing BAX- and BAK-mediated mitochondrial apoptosis. *Mol. Cell* **8**, 705–711 (2001).
15. Letai, A. *et al.* Distinct BH3 domains either sensitize or activate mitochondrial apoptosis, serving as prototype cancer therapeutics. *Cancer Cell* **2**, 183–192 (2002).
16. Veis, D. J., Sorenson, C. M., Shutter, J. R. & Korsmeyer, S. J. Bcl-2-deficient mice demonstrate fulminant lymphoid apoptosis, polycystic kidneys, and hypopigmented hair. *Cell* **75**, 229–240 (1993).
17. Motoyama, N. *et al.* Massive cell death of immature hematopoietic cells and neurons in Bcl-x-deficient mice. *Science* **267**, 1506–1510 (1995).
18. Ma, A. *et al.* Bclx regulates the survival of double-positive thymocytes. *Proc. Natl Acad. Sci. USA* **92**, 4763–4767 (1995).
19. Lee, P. P. *et al.* A critical role for Dnmt1 and DNA methylation in T cell development, function, and survival. *Immunity* **15**, 763–774 (2001).
20. Rickert, R. C., Roes, J. & Rajewsky, K. B lymphocyte-specific, Cre-mediated mutagenesis in mice. *Nucleic Acids Res.* **25**, 1317–1318 (1997).
21. Godfrey, D. I., Kennedy, J., Suda, T. & Zlotnik, A. A developmental pathway involving four phenotypically and functionally distinct subsets of CD3⁺CD4⁺CD8⁺ triple-negative adult mouse thymocytes defined by CD44 and CD25 expression. *J. Immunol.* **150**, 4244–4252 (1993).
22. Corcoran, A. E., Riddell, A., Krooshoop, D. & Venkataraman, A. R. Impaired immunoglobulin gene rearrangement in mice lacking the IL-7 receptor. *Nature* **391**, 904–907 (1998).
23. Hardy, R. R., Carmack, C. E., Shinton, S. A., Kemp, J. D. & Hayakawa, K. Resolution and characterization of pro-B and pre-pro-B cell stages in normal mouse bone marrow. *J. Exp. Med.* **173**, 1213–1225 (1991).
24. Shinkai, Y. *et al.* RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V(D)J rearrangement. *Cell* **68**, 855–867 (1992).
25. Kuhn, R., Schwenk, F., Aguet, M. & Rajewsky, K. Inducible gene targeting in mice. *Science* **269**, 1427–1429 (1995).
26. Bouillet, P. *et al.* Proapoptotic Bcl-2 relative Bim required for certain apoptotic responses, leukocyte homeostasis, and to preclude autoimmunity. *Science* **286**, 1735–1738 (1999).
27. Ranger, A. M. *et al.* Bad-deficient mice develop diffuse large B cell lymphoma. *Proc. Natl Acad. Sci. USA* **100**, 9324–9329 (2003).
28. Datta, S. R. *et al.* Survival factor-mediated BAD phosphorylation raises the mitochondrial threshold for apoptosis. *Dev. Cell* **3**, 631–643 (2002).
29. Akashi, K., Kondo, M., von Freuden-Jeffry, U., Murray, R. & Weissman, I. L. Bcl-2 rescues T lymphopoiesis in interleukin-7 receptor-deficient mice. *Cell* **89**, 1033–1041 (1997).
30. Maraskovsky, E. *et al.* Bcl-2 can rescue T lymphocyte development in interleukin-7 receptor-deficient mice but not in mutant *rag-1*^{-/-} mice. *Cell* **89**, 1011–1019 (1997).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank G. Zambetti for providing the anti-MCL-1 antisera, S. Dowdy for the Tat-Cre, T. Oltersdorf for the GST fusion constructs, J. Fisher and S. Wade for animal husbandry, and E. Smith for editorial assistance. J.T.O. is supported by a postdoctoral fellowship from the Damon Runyon Cancer Research Foundation. This work is supported in part by a grant from the National Institutes of Health.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.K. (stanley_korsmeyer@dfci.harvard.edu).

Optimization of specificity in a cellular protein interaction network by negative selection

Ali Zarrinpar^{1,2}, Sang-Hyun Park² & Wendell A. Lim²

¹Program in Biological Sciences, and ²Department of Cellular and Molecular Pharmacology, University of California, San Francisco, 600 16th Street, San Francisco, California 94143-2240, USA

Most proteins that participate in cellular signalling networks contain modular protein-interaction domains. Multiple versions of such domains are present within a given organism¹: the yeast proteome, for example, contains 27 different Src homology 3 (SH3) domains². This raises the potential problem of cross-

reaction. It is generally thought that isolated domain–ligand pairs lack sufficient information to encode biologically unique interactions, and that specificity is instead encoded by the context in which the interaction pairs are presented^{3,4}. Here we show that an isolated peptide ligand from the yeast protein Pbs2 recognizes its biological partner, the SH3 domain from Sho1, with near-absolute specificity—no other SH3 domain present in the yeast genome cross-reacts with the Pbs2 peptide, *in vivo* or *in vitro*. Such high specificity, however, is not observed in a set of non-yeast SH3 domains, and Pbs2 motif variants that cross-react with other SH3 domains confer a fitness defect, indicating that the Pbs2 motif might have been optimized to minimize interaction with competing domains specifically found in yeast. System-wide negative selection is a subtle but powerful evolutionary mechanism to optimize specificity within an interaction network composed of overlapping recognition elements.

How are SH3 domains used to assemble protein interaction networks with high specificity? One model postulates that domains have diverged sufficiently and have distinct recognition profiles (Fig. 1a). However, peptide library studies have shown that the recognition profiles of SH3 domains are highly overlapping^{5–7}: despite examples of domains with unusual recognition profiles⁸, most bind canonical proline-rich peptide motifs flanked by basic residues on either the amino terminus or the carboxy terminus (for

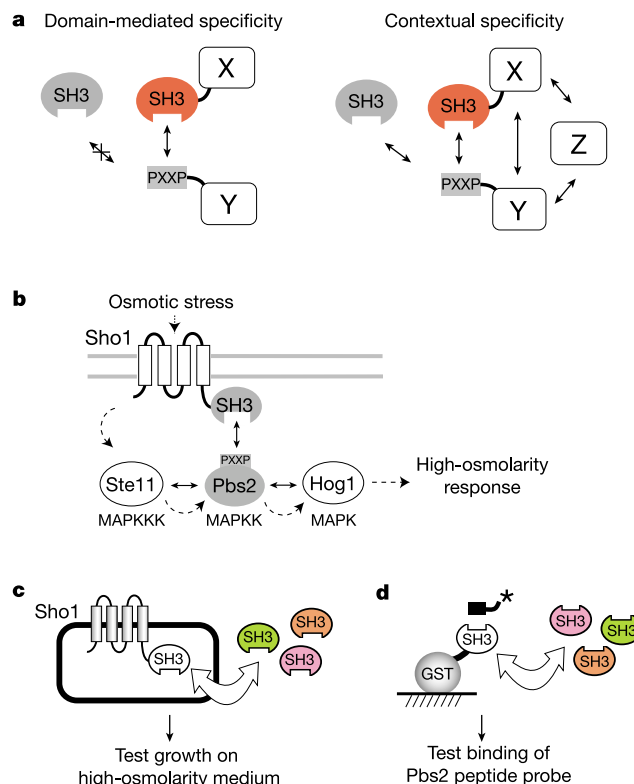


Figure 1 The yeast high-osmolarity pathway as a system for studying SH3 network specificity. **a**, Possible mechanisms of interaction specificity: domain-mediated specificity, in which individual domain–ligand pairs contain enough information to specify unique interactions, and contextual specificity, in which individual domain–ligand pairs lack sufficient information—factors such as cooperative interactions or subcellular localization are required to encode unique interactions. **b**, Interactions in the yeast high-osmolarity response MAP kinase (MAPK) pathway: solid arrows, physical interactions; dashed arrows, activating interactions. **c**, Growth assay to probe SH3 interchangeability *in vivo*: Sho1 chimeras bearing swapped SH3 domains were tested for ability to rescue the osmoresistance of *sho1Δ* strain. **d**, Array assay to test SH3 interchangeability *in vitro*: a set of GST–SH3 fusions arrayed on nitrocellulose were probed for binding to the tagged Pbs2–ligand peptide.

example, R/KXXPPXP or XPXXPR/K)^{9,10}. It is therefore postulated that specificity *in vivo* is encoded not by isolated domain-peptide partners but rather through the context in which the partners are presented (Fig. 1a)^{3,4}.

To examine the specificity of SH3 domain interactions, we probed the degree to which an SH3 domain from the yeast *Saccharomyces cerevisiae* could be interchanged with other SH3 domains. The fraction of alternative domains that cannot functionally replace the native domain reflects the interaction information content¹¹. If individual domains carry little specificity information, many SH3 domains should be able to act as functional replacements for a native SH3 domain.

The interaction of the SH3 domain from the yeast osmosensor protein Sho1 with a proline-rich motif from the kinase Pbs2 (Fig. 1b) is ideal for studying specificity. First, it is one of the few SH3 domain interactions that are unequivocally biologically relevant: it is essential for signalling in one branch of the high-osmolarity stress response pathway in yeast¹². Second, peptide library screens show that the Sho1 SH3 domain falls within the canonical SH3 recognition class (Supplementary Fig. S1)^{7,13}. Last, there are excellent methods for probing domain function. To probe specificity *in vivo*, we generated Sho1 chimaeras in which the wild-type domain was replaced by alternative SH3 domains (Supplementary Figs S2, S3) and tested their ability to rescue the growth of a Sho1 deletion strain on high-osmolarity media (Fig. 1c). To probe specificity *in vitro*, we made spatially defined arrays of SH3 domains fused to glutathione S-transferase (GST) and assayed these for binding to a labelled Pbs2 ligand (Fig. 1d).

Of 12 metazoan SH3 domains tested, six reconstituted osmo-

resistance when swapped into Sho1 (Fig. 2a). The same six domains bound to the Pbs2 ligand *in vitro* on SH3 domain arrays (Fig. 2b) and in solution binding assays using the free Pbs2 peptide (Fig. 2c). There was a good correlation between binding affinity and ability to rescue function (Supplementary Fig. S4), with a K_d of 40 μ M or less (wild-type $K_d = 1.3 \mu$ M) sufficient to restore detectable pathway function *in vivo*. These results are consistent with low information content in individual SH3 domains: the Pbs2 ligand motif is promiscuously recognized by this set of domains and the native domain is functionally interchangeable.

In contrast, a much higher level of specificity was observed when similar assays were performed with the set of 27 yeast SH3 domains. None of the 26 alternative SH3 domains could reconstitute osmo-resistance (Fig. 2d). This lack of function was not due to changes in expression or localization of chimaeric Sho1 protein (see Supplementary Fig. S5). Moreover, none of the alternative domains tested showed detectable binding to the Pbs2 peptide *in vitro* (Fig. 2e) on the arrays or by solution binding assays (Fig. 2f).

These results suggest that the isolated SH3 domain–ligand pair contains sufficient information to encode interaction specificity among the yeast set of SH3 domains. Several other observations support this model. A non-functional Sho1–SH3 chimaera can be complemented by compensatory changes in the Pbs2 peptide motif (Fig. 2g, Supplementary Fig. S6): the yeast Abp1 SH3 domain reconstitutes function only when combined with Pbs2 bearing an Abp1-binding peptide¹⁴. Moreover, the native interaction pair can be functionally replaced by a heterologous non-SH3 interaction—a PDZ domain interaction¹⁵ (Fig. 2g). Thus, diverse interactions can functionally replace the wild-type SH3-domain–ligand pair, as long

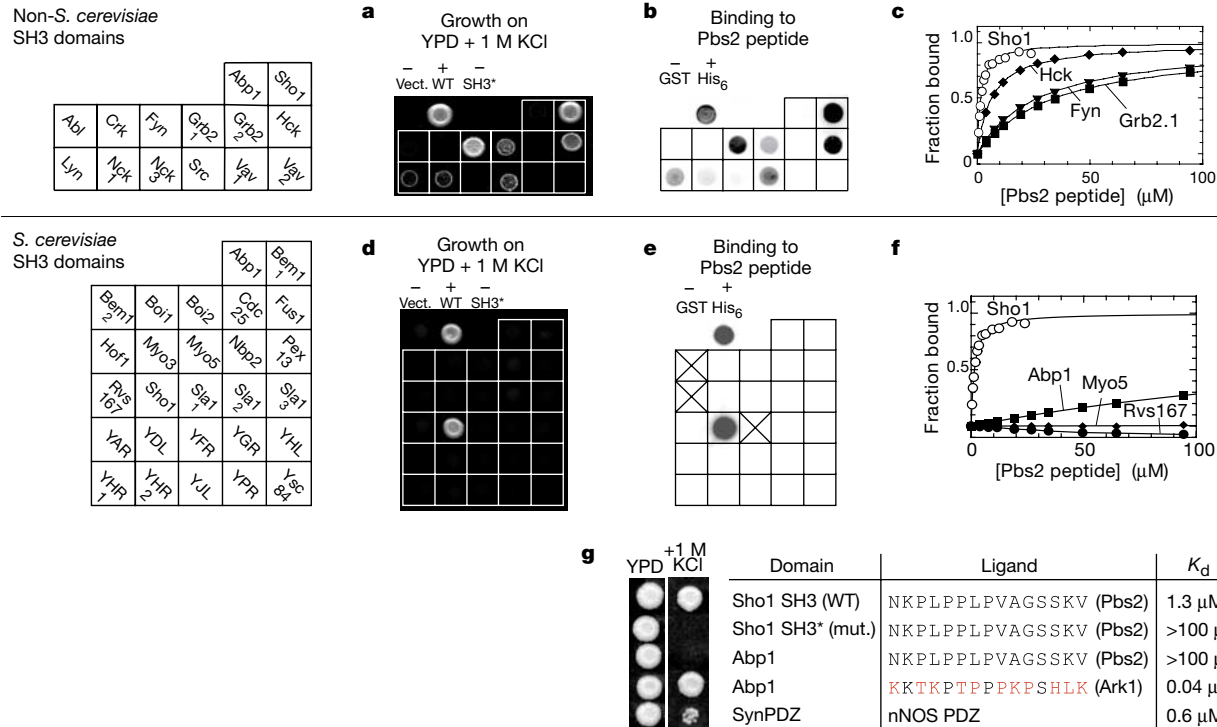


Figure 2 The Sho1 SH3 domain can only be functionally replaced by extra-species (non-*S. cerevisiae*) SH3 domains. **a, d**, Growth of cells containing Sho1 chimaeras (swapped SH3 domains) on high-osmolarity medium. Positive control is wild-type Sho1 (WT); negative controls are vector alone and Sho1 with a non-binding mutation (SH3*: W338F (ref. 30; A.Z., unpublished observations)). The arrangement is shown on the left; subscript indicates domain number in multidomain proteins. **b, e**, SH3 binding arrays. GST–SH3 fusion proteins spotted on nitrocellulose were probed with His-tagged Pbs2 peptide and anti-His antibody. The positive control was a His-tagged protein directly

spotted on membrane; the negative control was GST alone. Positions marked with X indicate insoluble SH3 fusions. **c, f**, Solution binding assays of Pbs2 peptide to representative SH3 domains. **g**, Compensatory changes in Pbs2 ligand can rescue the osmo-resistance of non-functional Sho1 chimaeras. A Pbs2 variant bearing a motif that binds Abp1 SH3 complements a Sho1–Abp1 chimaera. A heterologous PDZ–PDZ-domain interaction pair from syntrophin (Syn) and neuronal nitric oxide synthase (nNOS) can also reconstitute osmo-resistance.

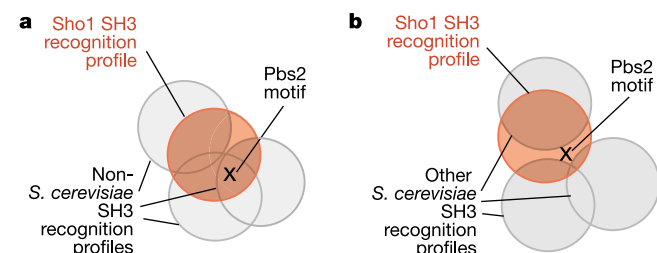


Figure 3 Model: role of proteome-wide negative selection in interaction network specificity. **a**, Non-*S. cerevisiae*: Pbs2 peptide is a canonical motif falling within the recognition space (circles) of several SH3 domains and therefore shows cross-reactivity with many non-yeast SH3 domains. **b**, *S. cerevisiae*: negative selection against cross-reactivity with physiologically relevant competitor domains (that is, other yeast domains) could drive the Pbs2 motif into a sequence space niche compatible only with the Sho1 SH3 domain (red circle).

as the interaction is of sufficient affinity. Other yeast SH3 domains cannot be functionally swapped into Sho1 because they simply do not cross-react with Pbs2.

Why does the interaction between the Sho1 SH3 domain and the Pbs2 ligand show such a high level of specificity within the set of yeast SH3 domains, but not within a set of non-yeast SH3 domains? There is no simple explanation based on sequence clustering of the two SH3 domain sets (Supplementary Figs S2, S3). Instead, an attractive model is that the specificity observed in yeast SH3 domains results not only from positive selection of the Pbs2 ligand

for interaction with the Sho1 SH3 domain, but also from negative selection against binding to competing SH3 domains from the same organism (Fig. 3). If the recognition profile of the Sho1 domain overlaps with those of many other SH3 domains, most random ligands that bind Sho1 will show high levels of cross-reactivity (Fig. 3a). However, if the Pbs2 motif were specifically selected to minimize cross-reaction with other yeast SH3 domains (Fig. 3b), specificity would be observed only within the yeast domain set and not within the non-yeast domain set—only domains within the same proteome would be evolutionarily relevant targets for negative selection. In summary, this model suggests that because interaction domains proliferate over the course of evolution, specificity can be enhanced by a combination of increasing divergence in the domain recognition profiles and pruning of cross-reactivity by negative selection. Binding interactions might be rendered orthogonal through evolution in much the same way that organisms within a single ecosystem speciate to exploit distinct niches¹⁶.

One way of testing the importance of negative selection in interaction network optimization is to probe the sequence space around the Pbs2 motif (Fig. 4a). This model would predict a loss of specificity as the motif drifts away from its optimum. With this aim we made a library of 19 of the 47 possible single-base-pair missense mutations of the Pbs2 motif, leaving the core prolines unchanged (Fig. 4b, Supplementary Fig. S7). We then assayed the specificity and affinity of this ligand library by using yeast SH3 arrays (Fig. 4b). We used the intensity of the Sho1 spot as an index of peptide affinity for the Sho1 domain (Supplementary Fig. S8). As an index of peptide specificity we divided the intensity of the Sho1 spot by the average

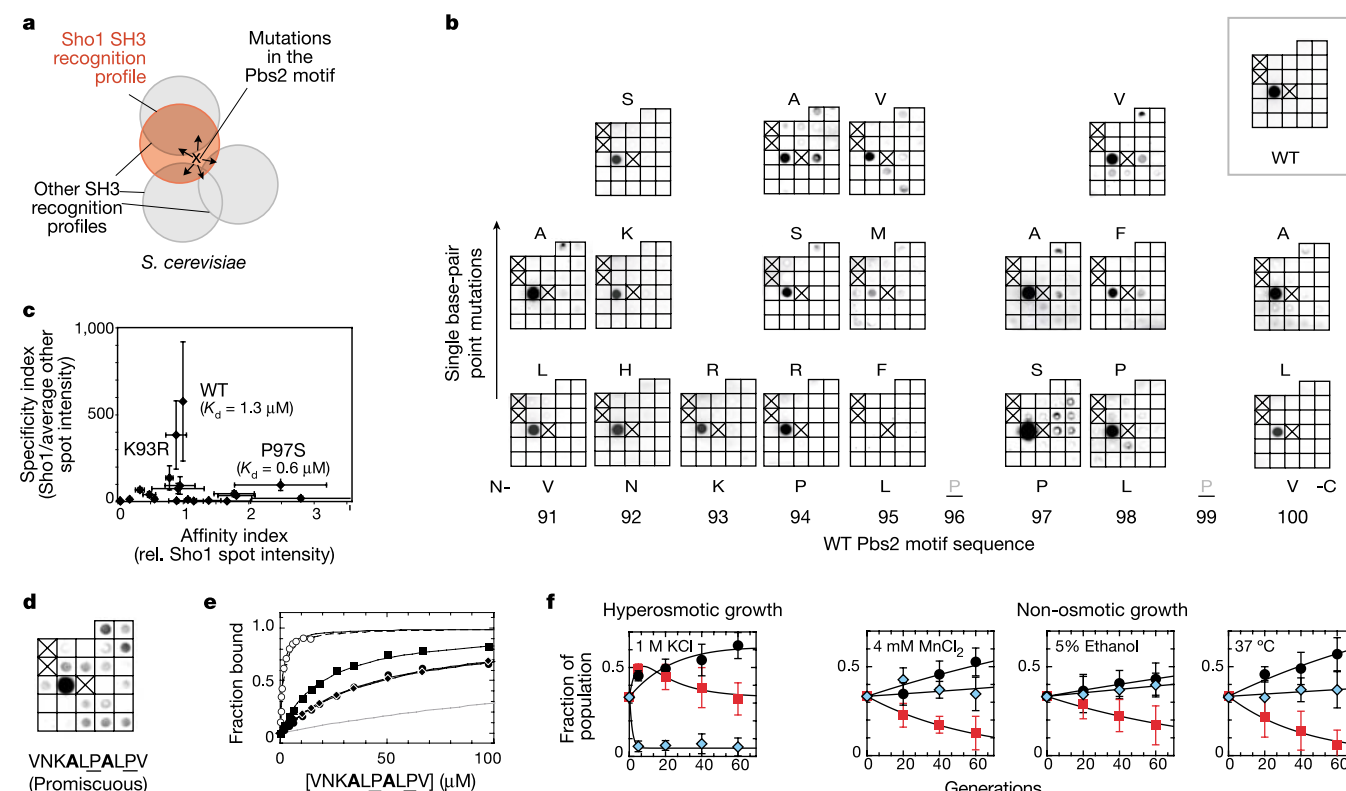


Figure 4 Pbs2 proline-rich motif is optimized to avoid cross-reactivity with other yeast SH3 domains. **a**, Mutational drift of Pbs2 motif is predicted to perturb network negative selectivity in *S. cerevisiae*. **b**, Yeast SH3 domain arrays probed for binding to Pbs2 peptide variants (19 of 47 possible missense point mutants; Supplementary Fig. S7). The wild-type (WT) sequence is shown at the bottom (conserved prolines underlined); point substitutions are indicated above. **c**, Quantification of mutant binding arrays (see Supplementary Fig. S8) shows that the wild-type Pbs2 motif is optimized for Sho1 interaction specificity but not for affinity. **d**, **e**, Combining P94A and P97A mutations yields

a highly promiscuous Pbs2 motif variant, assayed by array binding (**d**) and solution binding assays (**e**). Binding curves of wild-type peptide to Sho1 (dashed line) and Abp1 (grey line) SH3 domains are shown for comparison. Open circles, Sho1; squares, Abp1; filled circles, Rvs167; diamonds, Myo5. **f**, Competition growth assays starting with equal fractions of strains bearing wild-type Pbs2 (VNKLPPLPV; black circles), promiscuous Pbs2 (VNKALPALPV; red squares) and non-interacting Pbs2 (VNKLAALAV; blue diamonds) performed under various conditions.

intensity of the remaining 23 non-Sho1 spots. Some peptide mutations increased affinity, others decreased affinity; but all yielded an increase in cross-reactivity with other yeast SH3 domains (Fig. 4c). On the basis of this mutant analysis, several residues in the ligand seem to be more significant than others in determining specificity. However, it is difficult to rationalize these effects precisely on the basis of structural comparisons (Supplementary Fig. S9).

This analysis indicates that the wild-type Pbs2 motif is not optimized for affinity for the Sho1 SH3 domain. Rather, it seems to be optimized for specificity. In fact, by combining two point mutations (P94A, P97A), we constructed a promiscuous Pbs2 motif variant that bound to the Sho1 SH3 domain with slightly higher affinity than the wild-type Pbs2 motif, but with a markedly higher level of cross-reactivity (Fig. 4d, e). Thus the high specificity of the Sho1–Pbs2 interaction within the yeast SH3 interaction network is not the result of the Sho1 SH3 domain's having a distinct recognition profile; rather, it is the result of the ligand's exploiting niches in sequence space not recognized by other physiologically competitive SH3 domains.

The generation of Pbs2 variants that bind with high affinity to Sho1 but also cross-react significantly with other yeast SH3 domains affords us the opportunity to test the biological importance of interaction network specificity. We compared the fitness of strains containing different forms of Pbs2—wild-type, promiscuous mutant (P94A/P97A) or non-interacting mutant (P96A/P99A, core prolines)—under various growth conditions. Under hyperosmotic growth conditions, the non-interacting mutant strain, because it is osmosensitive, was rapidly overtaken by the other strains. The promiscuous mutant strain, in contrast, was competitive with the wild-type strain. However, under some conditions, such as growth in minimal medium at 37°C, the promiscuous mutant strain was overtaken by both the wild-type and non-interacting mutant strains (Fig. 4f). Thus, the promiscuous mutant strain seems to have a fitness defect under these conditions that is not due to a defect in the osmolarity response pathway. Promiscuous interactions might lead to small but evolutionarily important disadvantages.

The generality of negative selection as a mechanism for specificity enhancement is difficult to probe because so few biologically verified SH3-domain–ligand pairs in yeast have been clearly identified. Nevertheless, we examined two of the better-characterized yeast SH3 domains, those from Abp1 and Pex13 (Supplementary Fig. S10). A putative ligand for the Abp1 SH3 domain, a peptide from Ark1 (ref. 14), was observed to bind the Abp1 SH3 domain with minimal cross-reactivity against other yeast SH3 domains. In contrast, a proline-rich peptide from Pex14 was found to cross-react with seven other yeast SH3 domains in addition to the Pex13 domain, its native partner¹⁷ (Supplementary Fig. S10). However, previous findings have shown that a functional interaction of Pex13 and Pex14 is dependent on the interaction of both of these proteins with a third protein Pex5 (ref. 18), a case of multi-partner cooperativity in recognition. Moreover, cellular localization studies show that Pex13 is the only SH3-domain-containing protein in peroxisomes^{2,19}. Pex14 also localizes to the peroxisome independently of the Pex13 SH3 domain²⁰. In contrast, Sho1 and Pbs2 both overlap in subcellular localization with up to 16 other SH3-domain-containing proteins^{2,19}. Thus, because of subcellular colocalization and cooperative interactions, the Pex13–Pex14 interaction pair might not have had the same selective pressure to achieve the level of discrimination observed for Sho1–Pbs2. It is also possible that in some cases binding promiscuity is required for function²¹. These results show how negative selection is only one of several possible mechanisms used to enhance interaction specificity.

Thus, negative domain–ligand selection can have a powerful function in optimizing protein interaction network specificity. The power of negative selection has previously been recognized in

immunology²² and is likely to have a key function in the construction of many biological systems, including signalling²³ and transcriptional networks²⁴. The importance of negative selection suggests that to map cellular interaction networks it will be critical not only to search for potential ligands with optimal affinity but also to characterize cross-reactivity of these ligands with relevant sets of competing receptors. In higher eukaryotes, in which only a fraction of a genome is expressed in each cell type^{25,26}, accurate interaction mapping might require the characterization of cell-type-specific domain expression profiles to delineate physiologically competitive domain sets. □

Methods

Constructs and strains

Yeast strains (W303 background) were all derived from the *ssk2Δ,ssk22Δ* mutant, which lacks the Sho1-independent branch of the osmoregulation pathway¹³. Sho1 chimaeras were constructed as shown in Supplementary Fig. S2 and expressed from a CEN/ARS plasmid (pRS316) with the native Sho1 promoter (strain *ssk2Δ,ssk22Δ,sho1Δ* or *ssk2Δ,ssk22Δ,pbs2Δ,sho1Δ*). Pbs2 mutants were constructed in pRS304 with native Pbs2 promoter (Supplementary Fig. S6) and integrated as a single copy into the genome (strain *ssk2Δ,ssk22Δ,pbs2Δ* or *ssk2Δ,ssk22Δ,pbs2Δ,sho1Δ*).

Protein expression and purification

Sho1, Pbs2 and all *S. cerevisiae* SH3 domains were cloned by polymerase chain reaction (PCR) from genomic DNA. Metazoan SH3 domains were cloned from appropriate cDNA libraries. Fragments encoding the SH3 domains were ligated into a Sho1 yeast vector for growth assays (Supplementary Fig. S2) and a GST-fusion vector for bacterial expression. His₆-tagged Pbs2 peptides, fused to the N-terminal domain of lambda repressor (residues 1–99), were constructed as described²⁷ for use as probes for SH3 arrays. Recombinant proteins were expressed in *Escherichia coli* strain BL21 (DE3) RIL. His₆ fusions and GST fusions were purified on Ni²⁺-nitrilotriacetate resin (Qiagen) or glutathione agarose, respectively, by standard methods (elution with 250 mM imidazole or 10 mM glutathione, respectively). Protein concentrations were quantified by ultraviolet radiation absorbance.

Hyperosmotic plate growth assay

Cells (10³) were spotted on YPD plates with or without 1 M KCl (Supplementary Fig. S11). Plates were incubated at 30°C for 3–5 days.

Peptide synthesis

Peptides (acetylated and amidated) were synthesized with standard 9-fluorenylmethoxycarbonyl (Fmoc) chemistry and purified by reverse-phase chromatography. Molecular masses were verified by mass spectrometry. Concentrations were verified by quantitative amino acid analysis.

SH3 domain array binding assay

100 μl each of 0.1 μM purified GST–SH3 fusion proteins in TBST buffer were spotted in array format on pretreated nitrocellulose membrane with a dot-blot apparatus. Array membranes were blocked in 3% milk, 1% BSA in TBST for 1 h at room temperature, and then probed with 6 ml of the His-tagged fusion protein containing the proline-rich peptide of interest (50 μM) in TBST for 4–16 h at 4°C. The membrane was washed four times in TBST, then reprobed for 1 h with a horseradish peroxidase-conjugated anti-His6 antibody (dilution 1:2000; Santa Cruz Biotechnology) at 4°C. Finally, the blot was developed with an enhanced chemiluminescence system and quantified on an AlphaImnotech charge-coupled device camera and analytical software. To control for variation in antibody levels and development exposure, standards of a His-tagged protein (100 μl of 100-nM and 10-nM solutions) were directly spotted on the membrane.

Spot intensities were corrected for background and for variations in spotting and exposure (Supplementary Figs S8a, S11). Corrected intensities for each spot are given relative to the intensity for the Sho1 SH3-domain spot probed with wild-type peptide. The semiquantitative nature of this assay was validated by comparing spot intensities from the SH3-domain arrays with dissociation constants measured *in vitro* (Supplementary Fig. S8b).

Measurement of binding affinities

Affinities of synthetic peptide ligands for SH3 domains were measured by following the increase in domain tryptophan fluorescence on titration of ligand into a solution of SH3 domain at a fixed concentration of 0.01–0.5 μM (always less than 0.25K_d)²⁸. The ligand stock concentration was typically between 0.1 and 2 mM. Data were fitted to the equation

$$y = F_0 + \left[\frac{(F_{\max} - F_0) \cdot x}{K_d} \right] / \left[1 + \frac{x}{K_d} \right]$$

by nonlinear least-squares analysis with the program ProFit 5.6.4 (Quantum Soft), where y is the fluorescence reading, x is ligand concentration, K_d is dissociation constant, F_0 is initial fluorescence value (fraction bound = 0) and $F_{\max}$ is fluorescence value at saturation (fraction bound = 1).

Competition growths

Starter cultures of the three strains (Supplementary Fig. S6) were grown independently to an OD₆₀₀ of 0.5. Equal amounts of each were combined into one tube. An aliquot was

removed from this tube as a standard for comparison with subsequent time samples. Cells were diluted 1:100 into various media and incubated at the appropriate temperature until OD_{600} was 0.5, whereupon they were diluted 1:100 (about once or twice a day). Samples were removed at various time points and lysed by incubation with Zymolyase and boiling. The lysates were subjected to PCR. The PCR product was sequenced in accordance with the standard protocol provided by Applied Biosystems and analysed on an ABI Prism 3700 DNA Analyzer with DNA Sequencing Analysis Software Version 3.6.1 (Applied Biosystems, Foster City, California). Mutant frequencies within the culture were estimated with the sequencing-based protocol developed by Kwok and Duan²⁹. Data were fitted to exponential equations that accounted for changes in growth of the mutant of interest and changes in growth of competitors.

Received 7 August; accepted 23 October 2003; doi:10.1038/nature02178.

- Pawson, T. & Nash, P. Assembly of cell regulatory systems through protein interaction domains. *Science* **300**, 445–452 (2003).
- Saccharomyces Genome Database, (<http://www.yeastgenome.org/>).
- Mayer, B. J. SH3 domains: complexity in moderation. *J. Cell Sci.* **114**, 1253–1263 (2001).
- Ladbury, J. E. & Arold, S. Searching for specificity in SH domains. *Chem. Biol.* **7**, R3–R8 (2000).
- Sparks, A. B. *et al.* Distinct ligand preferences of Src homology 3 domains from Src, Yes, Abl, Cortactin, p53bp2, PLCgamma, Crk, and Grb2. *Proc. Natl Acad. Sci. USA* **93**, 1540–1544 (1996).
- Kay, B. K., Williamson, M. P. & Sudol, M. The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. *FASEB J.* **14**, 231–241 (2000).
- Cesareni, G., Panni, S., Nardelli, G. & Castagnoli, L. Can we infer peptide recognition specificity mediated by SH3 domains? *FEBS Lett.* **513**, 38–44 (2002).
- Zarrinpar, A., Bhattacharyya, R. P. & Lim, W. A. The structure and function of proline recognition domains. *Science* [online] (doi:10.1126/stke.2003.179.re8).
- Lim, W. A., Richards, F. M. & Fox, R. O. Structural determinants of peptide-binding orientation and of sequence specificity in SH3 domains. *Nature* **372**, 375–379 (1994).
- Feng, S., Chen, J. K., Yu, H., Simon, J. A. & Schreiber, S. L. Two binding orientations for peptides to the Src SH3 domain: development of a general model for SH3–ligand interactions. *Science* **266**, 1241–1247 (1994).
- Schneider, T. D. Evolution of biological information. *Nucleic Acids Res.* **28**, 2794–2799 (2000).
- Posas, F. & Saito, H. Osmotic activation of the HOG MAPK pathway via Ste11p MAPKKK: scaffold role of Pbs2p MAPKK. *Science* **276**, 1702–1705 (1997).
- Tong, A. H. *et al.* A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* **295**, 321–324 (2002).
- Fazi, B. *et al.* Unusual binding properties of the SH3 domain of the yeast actin-binding protein Abp1: structural and functional analysis. *J. Biol. Chem.* **277**, 5290–5298 (2002).
- Park, S. H., Zarrinpar, A. & Lim, W. A. Rewiring MAP kinase pathways using alternative scaffold assembly mechanisms. *Science* **299**, 1061–1064 (2003).
- Orr, M. R. & Smith, T. B. Ecology and speciation. *Trends Ecol. Evol.* **13**, 502–506 (1998).
- Barnett, P., Bottger, G., Klein, A. T., Tabak, H. F. & Distel, B. The peroxisomal membrane protein Pex13p shows a novel mode of SH3 interaction. *EMBO J.* **19**, 6382–6391 (2000).
- Bottger, G. *et al.* Saccharomyces cerevisiae PTS1 receptor Pex5p interacts with the SH3 domain of the peroxisomal membrane protein Pex13p in an unconventional, non-PXXP-related manner. *Mol. Biol. Cell* **11**, 3963–3976 (2000).
- Huh, W.-K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
- Girzalsky, W. *et al.* Involvement of Pex13p in Pex14p localization and peroxisomal targeting signal 2-dependent protein import into peroxisomes. *J. Cell Biol.* **144**, 1151–1162 (1999).
- Sudol, M. From Src Homology domains to other signaling modules: proposal of the 'protein recognition code'. *Oncogene* **17**, 1469–1474 (1998).
- Palmer, E. Negative selection—clearing out the bad apples from the T-cell repertoire. *Nature Rev. Immunol.* **3**, 383–391 (2003).
- Yaffe, M. B. *et al.* A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nature Biotechnol.* **19**, 348–353 (2001).
- Newman, J. R. & Keating, A. E. Comprehensive identification of human bZIP interactions with coiled-coil arrays. *Science* **300**, 2097–2101 (2003).
- Kim, S. K. *et al.* A gene expression map for *Caenorhabditis elegans*. *Science* **293**, 2087–2092 (2001).
- Jiang, M. *et al.* Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **98**, 218–223 (2001).
- Maxwell, K. L. & Davidson, A. R. Mutagenesis of a buried polar interaction in an SH3 domain: sequence conservation provides the best prediction of stability effects. *Biochemistry* **37**, 16172–16182 (1998).
- Lim, W. A., Fox, R. O. & Richards, F. M. Stability and peptide binding affinity of an SH3 domain from the *Caenorhabditis elegans* signaling protein Sem-5. *Protein Sci.* **3**, 1261–1266 (1994).
- Kwok, P. Y. & Duan, S. SNP discovery by direct DNA sequencing. *Methods Mol. Biol.* **212**, 71–84 (2003).
- Lim, W. A. & Richards, F. M. Critical residues in an SH3 domain from Sem-5 suggest a mechanism for proline-rich peptide recognition. *Nature Struct. Biol.* **1**, 221–225 (1994).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank A. Davidson and members of the Davidson laboratory, I. Herskowitz, P. Y. Kwok, E. O'Shea, A. Salí, K. Yamamoto, R. Zuckerman, P. Chien, K. Chak, R. Bhattacharyya, J. Dueber, N. Sallee, B. Yeh, and members of the Lim laboratory for assistance and discussion. This work was supported by grants from the NIH and the Packard Foundation. S.-H.P. is a Jane Coffin Child Fellow. A.Z. is supported by the UCSF Medical Scientist Training Program.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to W.A.L. (wlim@itsa.ucsf.edu).

Gating of the rapid shade-avoidance response by the circadian clock in plants

Michael G. Salter*, Keara A. Franklin* & Garry C. Whitelam

Department of Biology, University of Leicester, Leicester LE1 7RH, UK

* These authors contributed equally to this work

The phytochromes are a family of plant photoreceptor proteins that control several adaptive developmental strategies^{1,2}. For example, the phytochromes perceive far-red light (wavelengths between 700 and 800 nm) reflected or scattered from the leaves of nearby vegetation. This provides an early warning of potential shading, and triggers a series of 'shade-avoidance' responses, such as a rapid increase in elongation³, by which the plant attempts to overgrow its neighbours³. Other, less immediate, responses include accelerated flowering and early production of seeds. However, little is known about the molecular events that connect light perception with increased growth in shade avoidance. Here we show that the circadian clock gates this rapid shade-avoidance response. It is most apparent around dusk and is accompanied by altered expression of several genes. One of these rapidly responsive genes encodes a basic helix–loop–helix protein, PIL1, previously shown to interact with the clock protein TOC1 (ref. 4). Furthermore PIL1 and TOC1 are both required for the accelerated growth associated with the shade-avoidance response.

Selective absorption of blue and of red (600–700 nm) wavelengths by the chlorophylls means that the radiation reflected/scattered by green leaves is relatively enriched in the far-red (700–800 nm). This far-red-rich light signal (that is, a decrease in the ratio of red to far-red (R/FR)) is detected by nearby plants as a change in the equilibrium between the P_r and P_{fr} forms of phytochromes B, D and E (ref. 5), providing an unambiguous signal that potential competitors are nearby. In response to a low R/FR many plants evoke a suite of adaptive reactions, shade avoidance, including rapidly increased elongation of internodes⁶ and/or petioles, reduced leaf growth and increased apical dominance in an attempt to avoid being shaded. Prolonged exposure to the low-R/FR signal evokes a survival reaction: the acceleration of flowering^{3,7}. Shade avoidance is displayed by most angiosperms, including crop species, conferring high relative fitness in dense stands³ and is one of the best-studied examples of adaptive phenotypic plasticity in plants.

To gain insight into the molecular events involved in rapid shade-avoidance responses, we carried out Affymetrix *Arabidopsis* oligoarray analysis on plants exposed to low R/FR (see Supplementary Information). Among those genes displaying the most marked changes in expression in response to 1 h of low R/FR is the *ATHB-2* gene, encoding a homeodomain ZIP transcription factor; this gene is known to be rapidly and reversibly regulated by changes in R/FR (ref. 8). However, the greatest increase in transcript level in response to low R/FR was observed for a gene annotated as encoding an unknown protein. The transcript of this gene increases in abundance by ~35-fold at 1 h. After correcting for errors in the annotation of this gene, we identified it as *PIL1* (for *PIF3-like 1*) encoding a basic helix–loop–helix protein, previously identified as a protein that interacts with the circadian clock protein TOC1 (ref. 4).

The increase in *PIL1* transcript level in response to low R/FR starting 1 h after dawn is extremely rapid. Quantitative reverse transcriptase polymerase chain reaction (RT-PCR) shows that

increases in *PIL1* transcript levels are already detectable after only 8 min of low R/FR (Fig. 1a). Significantly, *PIL1* transcript levels fall very rapidly after transfer from low R/FR to high R/FR (data not shown). The Affymetrix oligoarray analysis indicated that *PIL1* transcript level was increased only about sixfold after 8 h of low R/FR treatment, but was again increased ~30-fold after 1 h of low R/FR the following dawn, a pattern suggestive of circadian regulation of *PIL1* expression. Quantitative RT-PCR analysis revealed a similar pattern of response in plants maintained for 72 h in continuous low-R/FR light, showing that regulation of *PIL1* by low R/FR was indeed under circadian control (Fig. 1b).

Because *PIL1* transcript levels did not oscillate in plants maintained under high R/FR, we investigated whether the circadian clock was gating the increase of *PIL1* transcript by low R/FR. Figure 1c shows that for plants grown under light/dark cycles and then transferred to continuous high R/FR, the increase in *PIL1* transcript in response to a 30-min reduction in R/FR shows a distinct and robust circadian oscillation. Brief decreases in R/FR at subjective dawn lead to more than an ~50-fold increase in *PIL1* transcript, whereas similar treatment at subjective dusk leads to about a 10–20-fold increase in *PIL1* transcript levels.

Expression of the *PIL1* gene has previously been reported to be repressed by exposure of etiolated seedlings to either white light⁹ or far-red light¹⁰. In fully de-etiolated seedlings grown under light/dark cycles, *PIL1* transcript levels are very low during the day and do not display a detectable increase during the night (data not shown) unless plants receive a brief exposure to far-red light before darkness. This indicates that one or more relatively stable *P_{fr}* species repress *PIL1* expression during the night and that low-R/FR treatments during the day lead to a derepression of *PIL1* expression.

Mutants deficient in particular phytochromes were studied to

identify the photoreceptors regulating *PIL1* transcript levels. Figure 2 shows that the low-R/FR-dependent increase in *PIL1* transcript levels is severely attenuated in the *phyB* mutant and essentially absent in the *phyBphyDphyE* triple mutant. Furthermore, *PIL1* transcript levels are constitutively elevated in mutants lacking *phyB*, with the highest level observed in the *phyBphyDphyE* triple mutant (Fig. 2). These observations are consistent with other physiological studies showing that responses to low R/FR are primarily mediated by *phyB*, with additional roles of *phyD* and *phyE* (ref. 5).

The clear circadian gating of the derepression of *PIL1* transcript levels by low R/FR led us to investigate whether shade-avoidance growth responses to low R/FR are also under clock control. We investigated the regulation of hypocotyl elongation in *Arabidopsis* seedlings, a sensitive assay system for shade-avoidance responses¹¹. Seedlings were grown under light/dark conditions for 4 days and then transferred to continuous high-R/FR light. At various times seedlings were exposed to low R/FR for 2 h, whereas controls remained in high-R/FR light. Hypocotyl lengths were determined 24 h after the low-R/FR treatment. Figure 3a shows that a single transient decrease in R/FR given at around subjective dusk is sufficient to induce a 25–35% increase in hypocotyl length. The effectiveness of the low-R/FR treatment oscillates with a period of ~24 h, and this oscillation persists for at least 3 days in continuous light. A transient decrease in R/FR at subjective dawn led to a small but significant inhibition of hypocotyl elongation compared with untreated control seedlings (Fig. 3a). This response presumably reflects oscillation in the levels and/or action of *phyA* because the growth inhibition is absent in *phyA* seedlings (Fig. 3a). The circadian clock has been shown to regulate the expression of apophytochrome genes^{12,13} and for *phyA* there is evidence for weak circadian cycling of *phyA* protein levels, peaking at subjective dawn¹⁴. As expected, constitutively elongated *phyB* mutant seedlings display an attenuated elongation growth response to transient decreases in R/FR that is also gated by the circadian clock (Fig. 3b).

The circadian-gated, rapid, reversible regulation of *PIL1* transcript by changes in R/FR and the circadian gating of rapid elongation growth responses to low R/FR occur with different phases. Maximal derepression of *PIL1* occurs at subjective dawn and maximal growth promotion occurs at subjective dusk. Nevertheless, even at subjective dusk there is a significant derepression of *PIL1* transcript levels, which led us to test whether *PIL1* expression is

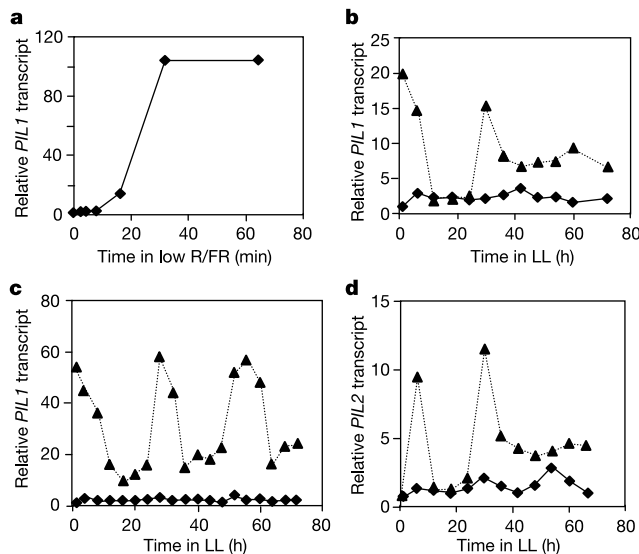


Figure 1 Quantitative PCR of *PIL1* transcript in response to low R/FR in *La-er*. **a**, Rapid time course showing the increase in *PIL1* transcript abundance after transfer to low R/FR. **b**, Time course of *PIL1* transcript abundance in plants transferred to continuous low R/FR, starting 1 h after dawn. Diamonds, high R/FR; triangles, low R/FR. **c**, Circadian gating of *PIL1* transcript abundance by low R/FR. Plants were either maintained in high R/FR (diamonds) or exposed transiently (30 min) to low R/FR (triangles). **d**, Time course of *PIL2* transcript abundance in plants transferred to continuous low R/FR. Diamonds, high R/FR; triangles, low R/FR. Each value is the mean of four separate quantitative PCR reactions normalized to *actin2*. Relative transcript abundance is normalized to *PIL1/PIL2* levels at time zero. Variation between individual reactions within treatments was less than 2%.

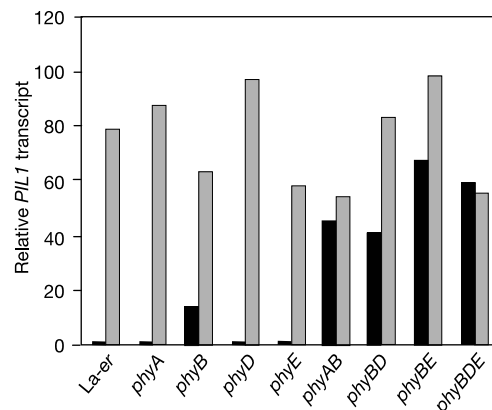


Figure 2 Quantitative PCR of changes in *PIL1* transcript abundance in *phy* mutants in response to low R/FR (grey bars), compared with control (black bars). Each value is the mean of three separate quantitative PCR reactions normalized to *actin2*. Relative transcript abundance is normalized to *PIL1* levels in untreated wild-type *La-er* plants. Variation between individual reactions within treatments was in all cases less than 2%.

required for growth responses to alterations in R/FR. We obtained a *pil1* knockout mutant (Garlic line 438c01) carrying a T-DNA insertion towards the 3' end of the coding region of the gene. Full-length *PIL1* transcript could not be detected in RNA samples from the mutant, and quantitative RT-PCR analysis also failed to detect any putative truncated transcript derived from the region that is 5' of the T-DNA (data not shown).

The mutant resembles wild-type seedlings after growth under a range of conditions (data not shown). As in the *La-er* ecotype, a single exposure of *pil1* knockout mutant (Col) seedlings to 2 h of low R/FR at subjective dusk is sufficient to induce a significant promotion of hypocotyl elongation, whereas a similar treatment at subjective dawn leads to an inhibition of elongation (Fig. 4a). The *pil1* seedlings display a marked attenuation of the maximal promotion of elongation by transient low R/FR (Fig. 4a) and an increase in the magnitude of the growth inhibition caused by transient low R/FR at subjective dawn. These data suggest that *PIL1* is required for the normal display of the shade-avoidance elongation responses to transient low R/FR. Furthermore, *pil1* seedlings exhibit a significant shift (~6 h) in the phase of the circadian rhythm of hypocotyl elongation responses to low R/FR. This phase shift seems to be specific for the circadian gating of the shade-avoidance elongation response, because it has been reported that other *pil1* knockout alleles display no obvious circadian phenotypes⁹.

We also generated *Ws*-ecotype seedlings that express the *PIL1* gene under the control of the strong 35S promoter (*PIL1ox*). Three independent homozygous lines were analysed in detail and all behaved in qualitatively the same manner. Treatment of the 35S:*PIL1* lines with low R/FR at dawn leads to a small additional increase in *PIL1* transcript levels, equivalent to that expected from photoregulation of the wild-type *PIL1* gene in these seedlings (data not shown). The shade-avoidance response of one of the lines (~100-fold higher *PIL1* transcript level than wild-type seedlings grown under high R/FR) is shown in Fig. 4b. In contrast with the *pil1* knockout, the inhibitions of elongation growth by low-R/FR treatments at subjective dawn are essentially abolished in the *PIL1ox* line (Fig. 4b), confirming that *PIL1* has a function in the regulation of hypocotyl elongation by R/FR. Significantly, the circadian rhythm of elongation growth response to low R/FR in the *PIL1ox* line is phase-shifted by ~6 h, as occurred in the *pil1* mutant.

The requirement for *PIL1* for the normal display of shade-avoidance responses allowed us to test whether the reported interaction between *PIL1* and *TOC1* (ref. 4) might be biologically relevant. For this, we investigated hypocotyl growth responses to low R/FR in the *toc1-2* mutant, in comparison with its wild-type

C24. Figure 4c shows that *TOC1* is also required for the normal promotion of elongation by low R/FR. At no point in the circadian cycle does a decrease in R/FR lead to a promotion of growth in the *toc1-2* mutant. The weak *toc1-1* allele¹⁵ also has an attenuated response to low R/FR (data not shown).

Previously, the reduced responsiveness of the strong *toc1-2* allele to both red and far-red light, with respect to inhibition of hypocotyl elongation and induction of gene expression, has implicated *TOC1* in phytochrome signalling¹⁵ in addition to its role in the circadian clock. We therefore investigated whether etiolated *pil1* mutant seedlings were defective in the light-mediated inhibition of hypocotyl elongation. Figure 4d and 4e shows that *pil1* seedlings are hyposensitive to both red and far-red light, particularly at lower fluence rates. These defective light responses resemble those reported for the *toc1-2* mutant¹⁵ and contrast with the report that other *pil1* knockout alleles do not have a defect in red-light-mediated inhibition of hypocotyl elongation⁹.

There have been many reports of circadian regulation of apo-

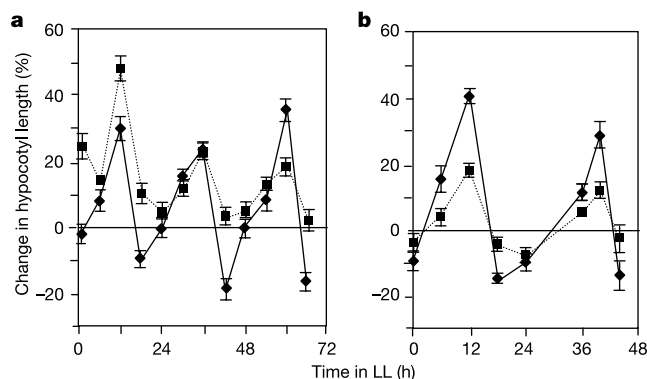


Figure 3 Circadian gating of low-R/FR-induced hypocotyl elongation in wild-type *La-er* (diamonds) and mutant (squares) seedlings. **a**, *phyA*; **b**, *phyB*. Treated plants were exposed to low-R/FR light for 2 h at different times before being moved back to white light for a further 24 h before harvesting and measurement. For each time point, the hypocotyls of at least 20 seedlings were measured.

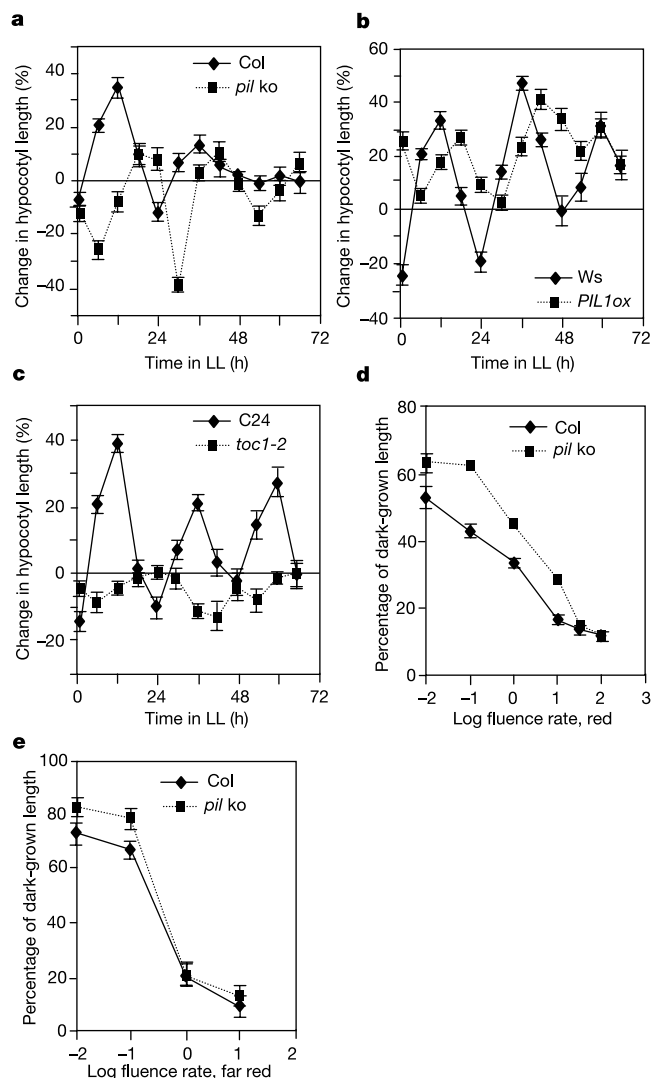


Figure 4 Hypocotyl elongation responses in *pil1*, *PIL1ox* and *toc1-2* seedlings. **a–c**, Circadian gating of low-R/FR-induced hypocotyl elongation in *pil1* (**a**), *PIL1ox* (**b**) and *toc1-2* (**c**) seedlings. Treated plants were exposed to low-R/FR for 2 h at different times before being moved back to white light for a further 24 h before harvesting and measurement. **d**, Hypocotyl elongation responses of *pil1* seedlings in continuous red (**d**) and far-red (**e**) light. Seeds were grown on Lele medium for 4 days at 22 °C. For each time point, the hypocotyls of at least 20 seedlings were measured.

phytochrome gene expression and of phytochrome-regulated responses, although most of these relate to early seedling de-etiolation and the expression of components of the photosynthetic machinery^{12,13,16}. The elongation of *Arabidopsis* hypocotyls is known to be under circadian control and is characterized by a daily arrest of elongation growth at dawn and a period of rapid elongation at dusk¹⁷. Our findings indicate that rapid elongation responses to low R/FR involve regulation of the rapid elongation phase of this rhythm. Furthermore, we have observed that phyA action might contribute to the dawn arrest of elongation growth, which is consistent with our earlier findings that *phyA* seedlings have elongated hypocotyls when grown in light/dark cycles¹⁸.

The rapidity of elongation growth responses to low R/FR has led to speculation that the mechanism might be independent of changes in gene expression¹⁹. However, our findings that the *PIL1* gene shows altered expression within a few minutes of a change in R/FR and that *PIL1* is necessary for the rapid elongation response indicate that this might not be the case. We propose that a change in P_{fr} level in phytochrome B, D and E leads to a rapid alteration in *PIL1* transcript and that consequent changes in *PIL1* protein level modulate the expression of target genes that are required for the differential elongation growth responses. Of course, our data indicate that although *PIL1* is necessary for the normal display of the rapid elongation response to low R/FR, an increase in *PIL1* transcript level is not sufficient for the response. This is most obvious from the finding that maximal derepression of *PIL1* expression by low R/FR occurs at subjective dawn, whereas maximal growth promotion induced by the same signal occurs at dusk when *PIL1* transcript levels increase only ~10–20-fold in response to the shade-avoidance signal. The observation that *TOC1* is also necessary for rapid shade avoidance might explain why the maximal growth promotion occurs at subjective dusk, a time when *TOC1* expression is greatest. It is possible that *TOC1* levels are limiting for the display of the elongation growth response to low R/FR. The expression of the *TOC1* gene does not respond to transient decreases in R/FR lasting for 2 h given at any time during the circadian cycle (data not shown).

We have also identified *PIL2*, a homologue of *PIL1*, that also shows circadian-gated increased transcript levels in response to low R/FR (Fig. 1d). Derepression of *PIL2* expression by low R/FR (a lag time of at least 3 h) is far less rapid than seen for *PIL1*. *PIL2* might be involved in longer-term responses to low R/FR, which is consistent with the observation that *pil1* mutant seedlings display more typical petiole elongation and flowering responses to prolonged exposure to low R/FR (data not shown). This suggests the operation of a *PIL1*-independent mechanism that can link perception of low R/FR with changes in elongation. □

Methods

Plant material and growth conditions

Expression analyses were performed with *La-er* wild-type plants grown to the 11-leaf stage in 12 h light/12 h dark cycles. Expression studies with *phy* mutants used seedlings that were grown in 12 h light/12 h dark cycles for 4 days and then given 30-min low R/FR starting 1 h after dawn. For each treatment 100 mg of seedlings were harvested and pooled. Hypocotyl experiments were performed with *phyA-1* (ref. 18; *La-er*), *phyB-1* (ref. 18; *La-er*) and *toc 1-2* (ref. 15; C24) mutant alleles and the corresponding wild-type plants. The *pil1* knockout mutant (*Col*) was obtained from Syngenta (Torrey Mesa Research Institute, San Diego, California). Garlic Line 438c01 and contains a T-DNA in the open reading frame. A homozygous line was generated with the herbicide resistance marker and segregation analysis along with PCR verification. Lines overexpressing *PIL1* (*PIL1ox*) were constructed in the *Ws* ecotype by subcloning the *PIL1* complementary DNA into the plant transformation vector CHF2 under the control of the constitutive 35S promoter. Transformed lines were selected on gentamycin.

For experiments on adult plants, *Arabidopsis* wild-type and mutant plants were germinated in 9-cm Petri dishes containing Lehle medium (Lehle Seeds Inc.) for 5 d in 12 h light/12 h dark conditions and then transferred to soil in trays. For circadian hypocotyl experiments, seeds were germinated on soil in 5-cm Petri dishes without lids, placed inside a moist chamber. All plants were grown in controlled-environment growth cabinets maintained at 22 °C. Low-R/FR experiments were performed with arrays of far-

red LEDs (λ_{max} 735 nm; Shinkoh Electronics). Plants maintained in high R/FR received a photon irradiance of 130 $\mu\text{mol m}^{-2} \text{s}^{-1}$ at 400–700 nm and a R/FR of 4.7. The low-R/FR treatments received the same photon irradiance but a R/FR of 0.089. For fluence-rate response curves, seedlings were sown on Lehle medium for 4 days at 22 °C. Red and far-red light were provided by LEDs. Measurement of hypocotyls was performed with ImageJ software (NIH).

Molecular biology methods

For Affymetrix arrays, RNA was initially extracted from the aerial tissue of six plants, ~11 true leaves, per treatment by using RNeasy midi kits (Qiagen) followed by extraction with acid phenol and precipitation with ethanol to a final concentration of >3 mg ml⁻¹. Hybridization to Affymetrix 8200 gene micro arrays, washing, staining and scanning were performed in accordance with the recommended protocol (Affymetrix, Santa Clara, California).

In all other experiments RNA extraction for analysis of transcript abundance used RNeasy mini preps (Qiagen) followed by treatment with DNasefree (Ambion) to remove contamination with DNA. cDNA pools were generated with an Omniscript RT-PCR kit (Qiagen) and 2 μg total RNA per reaction. The reaction volume was increased to 40 μl with all reagent quantities doubled. Transcript abundance of assayed genes was assessed by quantitative PCR with the Stratagene MX4000 as described in ref. 20.

Primer sequences for *PIL1* were 5'-AAATTGCTCTCAGCCATTCTGG-3' (forward) and 5'-TTCTAAGTTTGAGGCGGACGACG-3' (reverse). Primer sequences for *PIL2* were 5'-CACCACCATGGATGATACTCTTC-3' (forward) and 5'-TTCTTGCAAGAGGCCAAAGATCC-3' (reverse). The annealing temperature for *Actin2*, *PIL1* and *PIL2* assays was 58 °C. For *PIL1* and *PIL2*, a primer concentration of 200 nM and a final concentration of MgCl₂ of 2 mM proved optimal.

Received 4 July; accepted 17 October 2003; doi:10.1038/nature02174.

1. Quail, P. H. *et al.* Phytochromes: Photosensory perception and signal transduction. *Science* **268**, 675–680 (1995).
2. Kendrick, R. E. & Kronenberg, G. H. M. *Photomorphogenesis in Plants* 2nd edn (Kluwer Academic, Dordrecht, The Netherlands, 1999).
3. Smith, H. & Whitelam, G. C. The shade avoidance syndrome: multiple responses mediated by multiple phytochromes. *Plant Cell Environ.* **20**, 840–844 (1997).
4. Makino, S., Matsushika, A., Kojima, M., Yamashino, T. & Mizuno, T. The APR1/TOC1 quintet implicated in circadian rhythms of *Arabidopsis thaliana*. I. Characterization with APR1-overexpressing plants. *Plant Cell Physiol.* **43**, 58–69 (2002).
5. Franklin, K. A. *et al.* Phytochromes B, D and E act redundantly to control multiple physiological responses in *Arabidopsis*. *Plant Physiol.* **131**, 1340–1346 (2003).
6. Morgan, D. C., O'Brien, T. & Smith, H. Rapid photomodulation of stem extension in light-grown *Sinapis alba* L. studies on kinetics, site of perception and photoreceptor. *Planta* **150**, 95–101 (1980).
7. Cerdan, P. & Chory, J. Regulation of flowering time by light quality. *Nature* **423**, 881–885 (2003).
8. Carabelli, M., Morelli, G., Whitelam, G. & Ruberti, I. Twilight-zone and canopy shade induction of the *Athb-2* homeobox gene in green plants. *Proc. Natl Acad. Sci. USA* **93**, 3530–3535 (1996).
9. Yamashino, T. *et al.* A link between circadian-controlled bHLH factors and the APR1/TOC1 quintet in *Arabidopsis thaliana*. *Plant Cell Physiol.* **44**, 619–629 (2003).
10. Tepperman, J. M., Zhu, T., Chang, H.-S., Wang, X. & Quail, P. H. Multiple transcription-factor genes are early targets of phytochrome A signaling. *Proc. Natl Acad. Sci. USA* **98**, 9437–9442.
11. Botto, J. F. & Smith, H. Differential genetic variation in adaptive strategies to a common environmental signal in *Arabidopsis* accessions: phytochrome-mediated shade avoidance. *Plant Cell Environ.* **25**, 53–63 (2002).
12. Tóth, R. *et al.* Circadian clock-regulated expression of phytochrome and cryptochrome genes in *Arabidopsis*. *Plant Physiol.* **127**, 1607–1616 (2001).
13. Hall, A., Kozma-Bognár, L., Tóth, R., Nagy, F. & Millar, A. J. Conditional circadian regulation of *PHYTOCHROME A* gene expression. *Plant Physiol.* **127**, 1808–1818 (2001).
14. Sharrock, R. A. & Clack, T. Patterns of expression and normalized levels of the five *Arabidopsis* phytochromes. *Plant Physiol.* **130**, 442–456 (2002).
15. Mas, P., Alabadi, D., Yanovsky, M. J., Oyama, T. & Kay, S. A. Dual role of TOC1 in the control of circadian and photomorphogenic responses in *Arabidopsis*. *Plant Cell* **15**, 233–236 (2003).
16. Harmer, S. L. *et al.* Orchestrated transcription of key pathways in *Arabidopsis* by the circadian clock. *Science* **290**, 2110–2113 (2000).
17. Dowson-Day, M. J. & Millar, A. J. Circadian dysfunction causes aberrant hypocotyl elongation patterns in *Arabidopsis*. *Plant J.* **17**, 63–71 (1999).
18. Johnson, E., Bradley, J. M., Harber, N. H. & Whitelam, G. C. Photoresponses of light-grown *phyA* mutants of *Arabidopsis*. Phytochrome A is required for the perception of daylength extensions. *Plant Physiol.* **105**, 141–149 (1994).
19. Smith, H. Phytochromes and light signal perception by plants—an emerging synthesis. *Nature* **407**, 585–591 (2000).
20. Halliday, K. J., Salter, M. G., Thingnaes, E. & Whitelam, G. C. The phyB-controlled flowering pathway is temperature sensitive and is mediated by the floral integrator FT. *Plant J.* **33**, 875–885 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank E. Halligan, D. Lowes and J. Lunec for assistance with the Affymetrix microarrays, T. Ingles for assistance in creating transgenic lines, and A. Millar for providing *toc1* alleles. This work was supported by grants from the Biotechnology and Biological Research Council, UK.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.C.W. (gcw1@le.ac.uk)

Backtracking by single RNA polymerase molecules observed at near-base-pair resolution

Joshua W. Shaevitz^{1*}, Elio A. Abbondanzieri^{2*}, Robert Landick⁴ & Steven M. Block^{2,3}

¹Department of Physics, ²Department of Applied Physics, and ³Department of Biological Sciences, Stanford University, Stanford, California 94305, USA

⁴Department of Bacteriology, University of Wisconsin, Madison, Wisconsin 53706, USA

* These authors contributed equally to this work

Escherichia coli RNA polymerase (RNAP) synthesizes RNA with remarkable fidelity *in vivo*¹. Its low error rate may be achieved by means of a 'proofreading' mechanism comprised of two sequential events. The first event (backtracking) involves a transcriptionally upstream motion of RNAP through several base pairs, which carries the 3' end of the nascent RNA transcript away from the enzyme active site. The second event (endonucleolytic cleavage) occurs after a variable delay and results in the scission and release of the most recently incorporated ribonucleotides, freeing up the active site. Here, by combining ultrastable optical trapping apparatus with a novel two-bead assay to monitor transcriptional elongation with near-base-pair precision, we observed backtracking and recovery by single molecules of RNAP. Backtracking events (~5 bp) occurred infrequently at locations throughout the DNA template and were associated with pauses lasting 20 s to >30 min. Inosine triphosphate increased the frequency of backtracking pauses, whereas the accessory proteins GreA and GreB, which stimulate the cleavage of nascent RNA, decreased the duration of such pauses.

Recent studies have implicated the nucleolytic activity of RNA polymerase as part of a proofreading mechanism²⁻⁴, similar to that found in DNA polymerases⁵. A key feature of this proofreading mechanism is a short backtracking motion of the enzyme along the DNA template (directed upstream, opposite to the normal direction of transcriptional elongation). Similar rearward movements are thought to accompany the processes of transcriptional pausing⁶⁻⁸, arrest^{9,10}, and transcription-coupled DNA repair¹¹. During backtracking, the transcription bubble shifts and the DNA-RNA hybrid duplex remains in register, while the 3' end of the RNA transcript moves away from the active site, and may even protrude into the secondary channel (nucleotide entrance pore) of the enzyme^{6,7,9}, blocking the arrival of ribonucleoside triphosphates (NTPs). In its backtracked state, RNAP is able to cleave off and discard the most recently added base(s) by endonucleolysis, generating a fresh 3' end at the active site for subsequent polymerization onto the nascent RNA chain. In this fashion, short RNA segments carrying misincorporated bases can be replaced, leading to the correction of transcriptional errors (Fig. 1a). Accessory proteins have been identified that increase transcriptional fidelity by preferentially stimulating the cleavage of misincorporated nucleotides: GreA and GreB for *E. coli* RNA polymerase⁴ and SII/TFIIS for eukaryotic RNA polymerase II^{2,3}.

We studied transcription by RNAP at physiological nucleotide concentrations using a new single-molecule assay together with improved optical trapping instrumentation. In combination, these achieve subnanometre resolution along with extremely low positional drift. Our current system is capable of near-base-pair resolution in individual records of RNAP displacements, and achieves base-pair resolution (<0.3 nm) in averages of multiple records. During an experiment, two beads are optically trapped in buffer above a microscope coverglass by independently steered laser traps.

A recombinant derivative of *E. coli* RNAP is bound specifically via a biotin-avidin linkage to the smaller of two polystyrene beads, while the transcriptionally downstream end of the DNA template (or the upstream end, in the case of assisting forces) is bound to the larger bead via a digoxigenin-antibody linkage, forming a bead-RNAP-DNA-bead 'dumbbell' (Fig. 1b).

The tension in the DNA was kept nearly constant (8.4 ± 0.8 pN), for loads both opposing and assisting transcription, by feedback control of the position of the optical trap holding the larger bead. A force of this magnitude has a negligible effect on transcription rates, and is well below the stall force for RNAP¹². An opposing load was applied in all experiments, except where noted. Transcriptional elongation was observed by measuring the position of the smaller bead as the polymerase moved (Fig. 2a). We chose to make the trap holding the larger bead an order of magnitude stiffer than that holding the smaller bead so that all motion appeared in the latter (see Methods). None of the components of the assay were attached to the coverglass surface: this isolates the system from drift of the microscope stage relative to the objective and other optics, which represented a major source of low-frequency noise in previous single-molecule studies¹²⁻¹⁷. Measured drift rates during our experiments were typically below 5 nm h^{-1} (data not shown). We recorded the transcriptional motion of over 150 individual RNAP molecules at 1 mM NTPs moving on a DNA template derived from the *E. coli* *rpoB* gene sequence. As previously noted^{12,13,15,17}, RNAP activity consists of periods of continuous motion interrupted by distinct pauses of variable duration (Fig. 2a). The velocity during the continuous-motion phase averaged $\sim 15 \text{ bp s}^{-1}$, but varied among molecules, consistent with earlier reports^{12,13,15}.

Computer analysis of RNAP records identified transcriptional pauses ranging from 1 s (our detection threshold) to more than 30 min. Only intervals where transcriptional elongation ceased and subsequently recovered were scored as pauses. Pausing events could

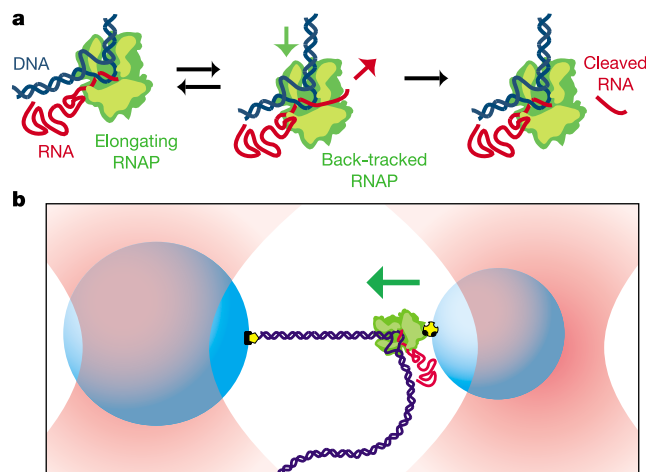


Figure 1 RNA polymerase transcription and proofreading studied by optical trapping. **a**, During normal elongation, RNAP (green) moves forward (downstream) on the DNA (blue) as it elongates the nascent RNA (red). At each position along the template, RNAP may slide backward along the template, causing transcription to cease temporarily. From the backtracked state, polymerase can either slide forward again, returning to its earlier state (left) or cleave the nascent RNA (right) and resume transcriptional elongation. **b**, Cartoon of the experimental geometry employed for opposing force experiments (not to scale). Two beads (blue) are held in separate optical traps (red) in a force-clamp arrangement. The smaller bead (right) is bound to a single molecule of RNAP, while the larger bead (left) is bound to the downstream end of the DNA by non-covalent linkages (yellow). During transcriptional elongation, the beads are pulled together. Nearly all the motion appears as a displacement of the right bead (green arrow), which is held in a comparatively weaker trap.

be broken up into two broad categories: 95% of events were 'short,' with lifetimes drawn from a double-exponential distribution with time constants of 1.5 s and 6.5 s, similar to our previous findings¹². The remaining 5% of events were 'long,' with lifetimes >20 s and a broad, non-exponential temporal distribution. Long pauses occurred at positions randomly distributed along the DNA template, rather than at stereotyped locations, and appeared to be sequence-independent within the resolution of these experiments. On average, long pauses occurred with a frequency of $0.95 \pm 0.21 \text{ kb}^{-1}$, a value that corresponds closely to ribonucleotide misincorporation rates during RNA synthesis *in vitro*¹, suggesting a possible role for such pauses in proofreading.

Operationally, we defined the duration of a 'pause' as the interval between the cessation of forward transcriptional motion and its subsequent recovery. At high spatial resolution, however, long pauses were found to consist of three distinct phases of motion that could be discerned in some individual records (Fig. 2b), as well as in averages of multiple records (see below). After abruptly stopping forward transcription, the enzyme underwent a slow rearward movement (phase 1, backtracking), typically lasting from 1–5 s, before stopping altogether for a variable interval (phase 2, pause). At the end of phase 2, rather than immediately resuming transcription at normal rates, RNAP moved forward gradually, typically for 3–10 s, transitioning to elongation mode only after a significant fraction of the initial backtracking distance had been retraced (phase 3, recovery). In contrast, neither the backtracking nor the recovery phases were evident in records of short pauses, where transitions both to and from normal elongation were abrupt (Fig. 2c).

To analyse the mean behaviour during phases 1 and 3, we averaged records of long pauses obtained under opposing loads after placing these in register along their rising edges, immediately before the cessation or the resumption of elongation, respectively (Fig. 3a). This procedure allows one to probe details of the motion that would otherwise be obscured by noise in individual records: similar trace-averaging techniques have been successfully employed to detect nanoscale steps in motor proteins such as myosin¹⁸ and NCD¹⁹, as well as to look for fast transients within the 8-nm step of kinesin²⁰. The average backtracking displacement during phase 1 of long pauses was $4.7 \pm 0.8 \text{ bp}$, and could be fitted by a decaying exponential with a time constant of $1.2 \pm 0.1 \text{ s}$. Both the duration and frequency of backtracking pauses are expected to display a strong force dependence due to the underlying motions involved. We found that the frequency of long pauses decreased dramatically from $0.95 \pm 0.21 \text{ kb}^{-1}$ under an $\sim 8 \text{ pN}$ opposing load to below 0.03 kb^{-1} under an $\sim 8 \text{ pN}$ assisting load (Table 1). This finding is consistent with a previous report of force dependence in the duration of very long pauses ($\sim 90 \text{ s}$) using a low-resolution optical trapping assay¹³.

Averages of records during phase 3 displayed a gradual forward motion, at an average velocity of $0.29 \pm 0.01 \text{ bp s}^{-1}$, before the resumption of normal elongation at $13.2 \pm 0.1 \text{ bp s}^{-1}$ (Fig. 3a). The average forward displacement during recovery was $2.5 \pm 1.0 \text{ bp}$, that is, about half of the initial backtracking distance. This reduced distance may reflect a mixed population of records, some of which exited from the pause more abruptly than others. However, the difference might also reflect the trace-alignment procedure. The exit from phase 2 is far less distinct than the entry into phase 1, and is therefore harder to pinpoint: minor registration errors tend to alter the magnitude of motions in averaged traces.

For comparison, we aligned and averaged an identical number of short pause records (Fig. 3d). In contrast to the long pause average, the short pause average displayed sharp transitions both into and out of the pause (that is, no phase 1 or phase 3 motions), with no associated movement greater than a base pair. The average of a much larger population of short pauses (>250) showed the same behaviour (data not shown). The absence of backtracking in short

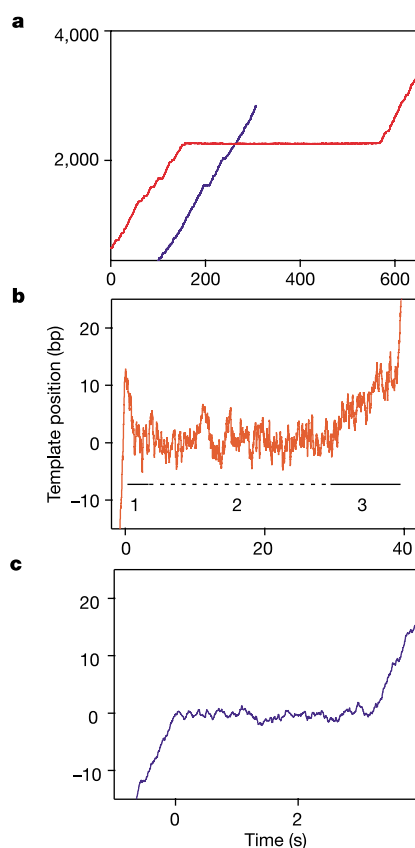


Figure 2 Backtracking occurs upon entry into long, but not short, pauses. **a**, Transcription records of two individual RNAP molecules are shown, each moving over the same template sequence. Both traces contain multiple short pauses (most are too short to be seen on this timescale); one includes a very long pause (410 s, red trace). **b**, In some records of long pauses, backtracking could be seen by eye: a representative record is shown. The three phases of motion are indicated below the trace: phase 1 (backtracking, solid line), phase 2 (pause, dotted line), and phase 3 (recovery, solid line). **c**, A representative record of a short pause (3 s); such pauses do not exhibit backtracking. Data were recorded at 2 kHz and boxcar-filtered at 100 ms for display.

pauses directly confirms and extends the conclusions of a recent study that examined the frequency and duration of short pauses, and found that these were independent of external load. This lack of force dependence implies an absence of backtracking motion, even as small as a single base pair¹².

We performed parallel experiments in the presence of the ribonucleotide analogue inosine triphosphate (ITP). Inosine mimics guanosine, forming a weak Watson–Crick pair that is slightly more stable than some measured mispairings^{21,22}. ITP incorporation inhibits next-nucleotide addition in human polymerase II by an amount similar to that of a mismatched base³. In elongating complexes, inosine incorporation decreases the stability of the RNA–DNA hybrid, changing the relative stability of the backtracked and non-backtracked states, and also decreases the stability of secondary structures in the nascent RNA formed behind the complex. The addition of 200 μM ITP to the standard transcription buffer (containing 1 mM levels of GTP, CTP, ATP and UTP) increased both the frequency and duration of long pauses (Fig. 3e, f; Table 1; Supplementary Information). Both phase 1 and phase 3 of long pauses were quantitatively similar to those observed in the absence of ITP (Fig. 3b). However, ITP did not affect either the frequency or the duration of short pauses, nor the average transcriptional velocity between pauses.

To assay the effects of transcript cleavage on long pauses, we added the *E. coli* transcription factors GreA and GreB. Addition of

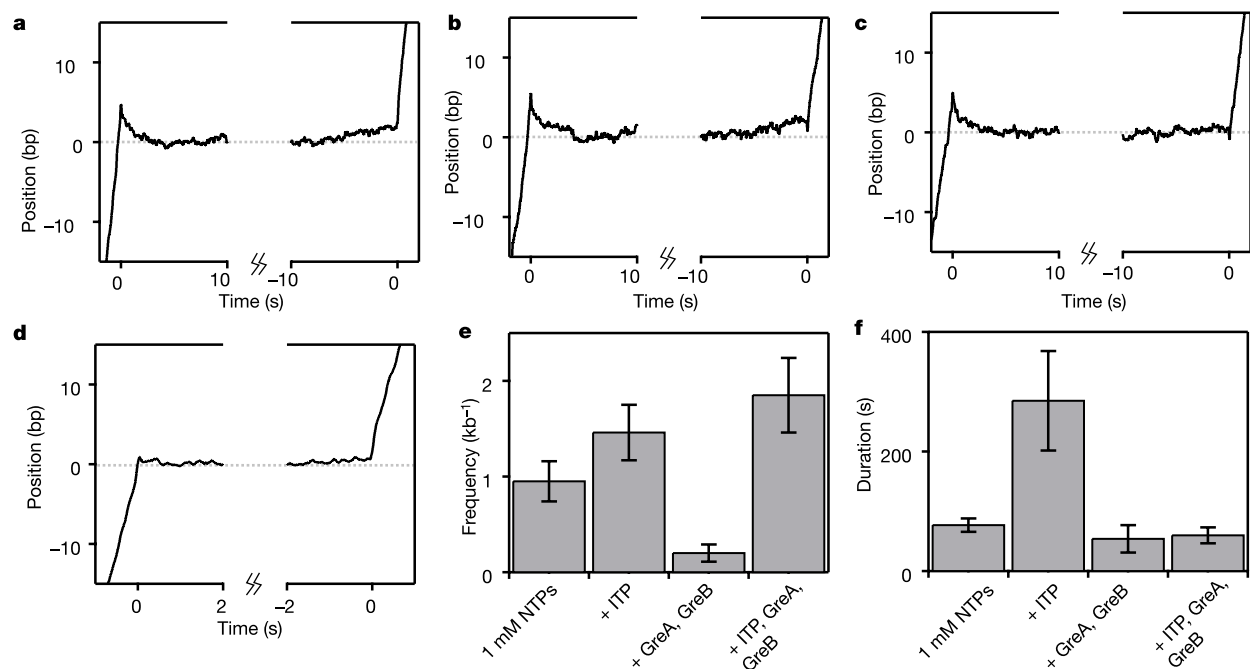


Figure 3 Averages of aligned long-pause records reveal details of backtracking and transcript cleavage events. **a**, Long-pause average of records at 1 mM NTPs ($N = 56$) displays a backtracking motion of ~ 5 bp (phase 1). Recovery (phase 3) is gradual, lasting ~ 5 s, before the resumption of normal elongation speed. **b**, Addition of ITP increases the frequency of long pauses that are indistinguishable from those in **a** ($N = 26$). **c**, Addition

of GreA and GreB reduced the duration of long pauses. Recovery from these pauses (phase 3) was abrupt ($N = 22$), distinct from **a** and **b**. **d**, The short pause average ($N = 56$) displays no backtracking motion. Average records were smoothed with a 100-ms boxcar filter for display. ITP and the accessory proteins GreA and GreB affect both the frequency (**e**) and duration (**f**) of long pauses. See text.

either 2 μ M GreA or 1 μ M GreB decreased the frequency of long pauses. Moreover, GreB significantly decreased the duration of long pauses, whereas the effect of GreA on duration was negligible (Fig. 3e, f; Table 1). The average distance backtracked in the presence of GreA increased to 6.6 ± 0.7 bp, whereas GreB appeared to abolish backtracking pauses altogether, yielding a mean backtrack distance close to zero for the few remaining long pauses, 0.5 ± 0.8 bp (Supplementary Information). These findings are consistent with the known properties of the two factors. GreA stimulates the cleavage of short, dinucleotide segments of backtracked RNA, and so its addition should relieve only those pauses associated with short backtracking motions, leaving pauses involving larger displacements. Conversely, GreB accelerates the removal of larger fragments to such a degree that the duration of backtracking pauses falls below the discrimination threshold of 20 s²³.

In the presence of both GreA and GreB, the duration (phase 2) of ITP-induced long pauses decreased dramatically, from 285 ± 83 s to 60 ± 13 s, while the average backtracking distance (phase 1) remained the same, 5.0 ± 1.0 bp (Table 1; Fig. 3f). Significantly, in the presence of both these transcription factors, the pause recovery (phase 3) became abrupt, and not gradual. We interpret this difference as being caused by Gre-stimulated cleavage of the RNA

blocking the secondary channel after backtracking. Such cleavage would lead to the prompt removal of oligonucleotides containing a potential mismatch, thereby reducing the lifetime of the backtracked, paused state (phase 2), and restoring the new 3' end of the nascent RNA to a position adjacent to the enzyme active site, ready for immediate polymerization.

In the absence of Gre-stimulated cleavage, RNAP must recover during phase 3 in a less direct fashion. In one mechanism, RNAP relies on its slow endogenous endonucleolytic activity to cleave the RNA at the enzyme active site, leading to long pause lifetimes but still allowing the possibility of error correction. In an alternative mechanism, thermal motions within the stalled complex may reverse the backtracking motion in a random walk process, carrying the original 3' end of the RNA back to the enzyme active site, once again leading to longer lifetimes, but without concomitant error correction. The gradual recovery seen during phase 3 in the absence of Gre factors may reflect these processes. Random fluctuations of polymerase motion during phase 2 would not be apparent in averaged records, which only show ensemble behaviour.

Taken together, our high-resolution, single-molecule experiments are consistent with a proofreading mechanism in *E. coli* RNA polymerase involving entry into an initial backtracked state of

Table 1 Long pause statistics under different experimental conditions

Condition*	Mean pause frequency (kb ⁻¹)	Mean pause duration (s)	Mean backtrack distance (bp)
1 mM ATP, UTP, GTP, CTP	0.95 ± 0.21 (<i>N</i> = 56)	77 ± 11	4.7 ± 0.8
1 mM NTPs, assisting force	<0.03 (<i>N</i> = 1)	—	—
+200 μM ITP	1.46 ± 0.29 (<i>N</i> = 26)	285 ± 83	5.5 ± 1.1
+2 μM GreA	0.14 ± 0.08 (<i>N</i> = 3)	56 ± 12	6.6 ± 0.7
+1 μM GreB	0.24 ± 0.09 (<i>N</i> = 8)	36 ± 7	0.5 ± 0.8
+2 μM GreA, 1 μM GreB	0.20 ± 0.09 (<i>N</i> = 5)	54 ± 23	5.8 ± 0.8
+200 μM ITP, 2 μM GreA, 1 μM GreB	1.85 ± 0.39 (<i>N</i> = 22)	60 ± 13	5.0 ± 1.0
+200 μM ITP, assisting force	0.15 ± 0.10 (<i>N</i> = 2)	—	—

* The applied force opposes transcription, unless otherwise noted.

the enzyme on the DNA template, followed by cleavage of the most recently polymerized RNA (1–10 bp) and enzymatic recovery. Under the conditions explored here (including an opposing load of ~8 pN), RNAP appeared to enter into long, backtracking pauses spontaneously at a rate of roughly once per kilobase: this rate was sensitive to transcription errors, and enhanced at least twofold by the addition of a nucleotide analogue. Incorporation of inosine leading to the backtracked state was relieved quantitatively by the action of transcription factors GreA and GreB, which are known to stimulate transcript cleavage. This simple editing mechanism may function, in principle, in many polymerase systems, including both prokaryotes and eukaryotes. □

Methods

Transcription assays

A bead–RNAP–DNA–bead dumbbell was constructed by binding a small 0.5- μ m diameter polystyrene bead to a biotin tag located on the β' subunit of a stalled *E. coli* RNAP transcription elongation complex, and a larger 0.7- μ m-diameter bead to the downstream end of the DNA template using a digoxigenin antibody (Fig. 1b). Stalled complexes and avidin-coated 0.5- μ m-diameter polystyrene beads were prepared as described previously¹². Polyclonal anti-digoxigenin antibody was covalently attached to carboxylated 0.7- μ m diameter polystyrene beads (Bangs Labs) via an EDC/Sulfo–NHS-coupled reaction. RNAP was stalled 29 base pairs after the T7A1 promoter on a template derived from the *rpoB* gene of *E. coli*¹⁶. Each of the two beads of the dumbbell was held in a separate optical trap ~1 μ m above the coverglass surface. Transcription along the DNA template was recorded by monitoring the position of the smaller bead (see below). All experiments were performed in transcription buffer (50 mM HEPES, pH 8.0, 130 mM KCl, 4 mM MgCl₂, 0.1 mM EDTA, 0.1 mM DTT, 20 mg ml⁻¹ heparin) in the presence of 1 mM NTPs and an oxygen scavenging system²⁴ at 22 $\pm$ 5 °C. ITP was purchased from Sigma-Aldrich Company. GreA and GreB proteins were purified as described²⁵.

Optical trapping

The salient aspects of the instrument used in these experiments have been described previously²⁶. Briefly, the apparatus is based on an inverted microscope (Nikon) modified for exceptional mechanical stability and the incorporation of two lasers for trapping and position detection. To form two optical traps, the trapping laser was split into two orthogonally polarized beams that can be steered independently. Software written in LabView 6i (National Instruments) controlled the position of the trapped 0.7- μ m-diameter bead via acousto-optical deflectors (IntraAction). The 0.5- μ m-diameter bead was illuminated by the detection laser: scattered light was projected onto a position-sensitive detector (Pacific Silicon Sensors) placed in a plane conjugate to the back focal plane of the microscope condenser. Bead position signals were smoothed at 1 kHz by an eight-pole low-pass Bessel filter (Krohn-Hite) and digitized at 2 kHz.

During an experiment under opposing loads, transcriptional elongation by RNAP shortens the DNA tether and leads to an increase in tension between the two beads. The laser power in each trap was high to reduce brownian noise in position, which is inversely proportional to the trap stiffness²⁷. The relative power in the two traps was fixed so that their ratio of stiffnesses was at least 1:10, which ensured that the majority of RNAP motion appeared as a change in the position of the smaller bead, held in the weaker trap. The resting tension in the DNA was maintained at 8.4 $\pm$ 0.8 pN (mean $\pm$ s.d.) by moving the 0.7- μ m-diameter bead in discrete 50-nm increments whenever the tension on the DNA exceeded 10 pN.

In our measurements, zero-mean, brownian fluctuations of the smaller bead represent the dominant source of noise. Position records of pauses along short pieces of DNA, that is, from enzymes that have already transcribed a substantial portion of the template, exhibit reduced noise because the amplitude of the brownian fluctuations is inversely proportional to the compliance of the linkage holding the bead. Additional sources of noise include tiny departures from sphericity in the beads (slightly non-spherical beads generate low-frequency noise through brownian rotation) and submicroscopic particle contaminants in the buffer that may fall into the optical trap (modulating the level of scattered light and thereby altering the apparent position). The combined effect of such sources causes the root mean square (r.m.s.) noise level to vary slightly from trace to trace: typically, noise levels were $\pm$ 5 bp at a 1-kHz bandwidth.

Data analysis

The contour length of the downstream DNA was computed from the measured position of the 0.5- μ m-diameter bead and the series elastic compliances of the optical traps and the DNA (using the nonlinear modified Marko–Siggia force–extension relation²⁸). Template position was determined by subtracting the initial tether length (4,226 bp for experiments using forces opposed to transcriptional elongation, and 1,406 bp for experiments using forces assisting transcriptional elongation) from the computed DNA contour length. We estimate our absolute error in determining the template position at $\pm$ 90 bp, due mainly to minor variations among individual bead diameters.

Transcriptional pauses in individual recordings were scored by eye and also by a custom computer algorithm (similar to ref. 12, but combining the numerical derivative and filtering operations into a single step). This algorithm recovers >99% of all pauses with lifetimes more than 1.5 s. Pauses occurred with a wide range of lifetimes, from seconds to many minutes. Broadly, events could be placed into two categories: 95% of all events belonged to a population that was satisfactorily fitted by a double-exponential relation with

time constants of 1.5 s (65% of the normalized amplitude) and 6.5 s (35% of the normalized amplitude). The remaining 5% of all events could not be so represented, and belonged to a much longer-lived population with a broader, non-exponential distribution. To better separate these populations for statistical analysis, we operationally defined 'short' pauses as having durations shorter than 5 s and 'long' pauses as having durations longer than 20 s, thereby excluding from analysis all pauses with lifetimes 5 < *t* < 20. Individual pause records were aligned along their initial and final rising edges^{18–20} and averaged to reduce measurement noise. Analysis was performed using Igor Pro 4.0 (Wavemetrics). All errors are reported as (mean $\pm$ s.e.m.), except for backtracking distances and durations, which were estimated using a bootstrap resampling analysis, and represent 68% confidence intervals.

Received 14 October; accepted 5 November 2003; doi:10.1038/nature02191.

Published online 23 November 2003.

- Erie, D. A., Yager, T. D. & von Hippel, P. H. The single-nucleotide addition cycle in transcription: a biophysical and biochemical perspective. *Annu. Rev. Biophys. Biomol. Struct.* **21**, 379–415 (1992).
- Jeon, C. & Agarwal, K. Fidelity of RNA polymerase II transcription controlled by elongation factor TFIIS. *Proc. Natl Acad. Sci. USA* **93**, 13677–13682 (1996).
- Thomas, M. J., Platas, A. A. & Hawley, D. K. Transcriptional fidelity and proofreading by RNA polymerase II. *Cell* **93**, 627–637 (1998).
- Erie, D. A., Hajiseyediavadi, O., Young, M. C. & von Hippel, P. H. Multiple RNA polymerase conformations and GreA: control of the fidelity of transcription. *Science* **262**, 867–873 (1993).
- Kunkel, T. A. & Bebenek, K. DNA replication fidelity. *Annu. Rev. Biochem.* **69**, 497–529 (2000).
- Komissarova, N. & Kashlev, M. RNA polymerase switches between inactivated and activated states by translocating back and forth along the DNA and the RNA. *J. Biol. Chem.* **272**, 15329–15338 (1997).
- Nudler, E., Mustaev, A., Lukhtanov, E. & Goldfarb, A. The RNA–DNA hybrid maintains the register of transcription by preventing backtracking of RNA polymerase. *Cell* **89**, 33–41 (1997).
- Marr, M. T. & Roberts, J. W. Function of transcription cleavage factors GreA and GreB at a regulatory pause site. *Mol. Cell* **6**, 1275–1285 (2000).
- Komissarova, N. & Kashlev, M. Transcriptional arrest: *Escherichia coli* RNA polymerase translocates backward, leaving the 3' end of the RNA intact and extruded. *Proc. Natl Acad. Sci. USA* **94**, 1755–1760 (1997).
- Reeder, T. C. & Hawley, D. K. Promoter proximal sequences modulate RNA polymerase II elongation by a novel mechanism. *Cell* **87**, 767–777 (1996).
- Tornaletti, S., Reines, D. & Hanawalt, P. C. Structural characterization of RNA polymerase II complexes arrested by a cyclobutane pyrimidine dimer in the transcribed strand of template DNA. *J. Biol. Chem.* **274**, 24124–24130 (1999).
- Neuman, K. C., Abbondanzieri, E. A., Landick, R., Gelles, J. & Block, S. M. Ubiquitous transcriptional pausing is independent of RNA polymerase backtracking. *Cell* **115**, 437–447 (2003).
- Forde, N. R., Izahy, D., Woodcock, G. R., Wuite, G. J. & Bustamante, C. Using mechanical force to probe the mechanism of pausing and arrest during continuous elongation by *Escherichia coli* RNA polymerase. *Proc. Natl Acad. Sci. USA* **99**, 11682–11687 (2002).
- Schafer, D. A., Gelles, J., Sheetz, M. P. & Landick, R. Transcription by single molecules of RNA polymerase observed by light microscopy. *Nature* **352**, 444–448 (1991).
- Wang, M. D. *et al.* Force and velocity measured for single molecules of RNA polymerase. *Science* **282**, 902–907 (1998).
- Yin, H., Landick, R. & Gelles, J. Tethered particle motion method for studying transcript elongation by a single RNA polymerase molecule. *Biophys. J.* **67**, 2468–2478 (1994).
- Adelman, K. *et al.* Single molecule analysis of RNA polymerase elongation reveals uniform kinetic behavior. *Proc. Natl Acad. Sci. USA* **99**, 13538–13543 (2002).
- Veigel, C. *et al.* The motor protein myosin-I produces its working stroke in two steps. *Nature* **398**, 530–533 (1999).
- deCastro, M. J., Fondecave, R. M., Clarke, L. A., Schmidt, C. F. & Stewart, R. J. Working strokes by single molecules of the kinesin-related microtubule motor ncd. *Nature Cell Biol.* **2**, 724–729 (2000).
- Nishiyama, M., Muto, E., Inoue, Y., Yanagida, T. & Higuchi, H. Substeps within the 8-nm step of the ATPase cycle of single kinesin molecules. *Nature Cell Biol.* **3**, 425–428 (2001).
- Aboul-ela, F., Koh, D., Tinoco, I. Jr & Martin, F. H. Base-base mismatches. Thermodynamics of double helix formation for dCA3XA3G+dCT3YT3G (X, Y=A,C,G,T). *Nucleic Acids Res.* **13**, 4811–4824 (1985).
- Martin, F. H., Castro, M. M., Aboul-ela, F. & Tinoco, I. Jr Base pairing involving deoxyinosine: implications for probe design. *Nucleic Acids Res.* **13**, 8927–8938 (1985).
- Borukhov, S., Sagitov, V. & Goldfarb, A. Transcript cleavage factors from *E. coli*. *Cell* **72**, 459–466 (1993).
- Yildiz, A. *et al.* Myosin V walks hand-over-hand: single fluorophore imaging with 1.5-nm localization. *Science* **300**, 2061–2065 (2003).
- Feng, G. H., Lee, D. N., Wang, D., Chan, C. L. & Landick, R. GreA-induced transcript cleavage in transcription complexes containing *Escherichia coli* RNA polymerase is controlled by multiple factors, including nascent transcript location and structure. *J. Biol. Chem.* **269**, 22282–22294 (1994).
- Lang, M. J., Asbury, C. L., Shaevitz, J. W. & Block, S. M. An automated two-dimensional optical force clamp for single molecule studies. *Biophys. J.* **83**, 491–501 (2002).
- Svoboda, K. & Block, S. M. Biological applications of optical forces. *Annu. Rev. Biophys. Biomol. Struct.* **23**, 247–285 (1994).
- Wang, M. D., Yin, H., Landick, R., Gelles, J. & Block, S. M. Stretching DNA with optical tweezers. *Biophys. J.* **72**, 1335–1346 (1997).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We acknowledge intellectual contributions from J. Gelles, and we thank the entire Block Laboratory, especially K. Neuman, for support and discussions. We also thank A. Meyer for reading of the original manuscript. This work was supported by grants from the NIGMS.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.M.B. (sblock@stanford.edu).

Spectroscopy and fluorescence

Flames, formats and photometry.

Photometer 8000

Palintest

www.palintest.com

Photometry made easy

The Palintest Photometer 8000 is designed for convenient use in labs carrying out water quality testing. Covering the full range of tests for clean and wastewater, it is compatible with all Palintest tablet and liquid photometer tests. With no moving parts, all measurements are made instantly. Compatible Palintest Tubetests reagent tubes are coded to provide all the instrument needs for wavelength range and calibration set-up and analysis. The instrument recognizes the tube method and measurement range and automatically sets up for a measurement. The multi-size cell holder adjusts automatically to all reagent tubes from 13 mm to 20 mm diameter with no inserts or adaptors required. The system memory allows up to 1000 results to be stored on-board and selectively or batch uploaded to LIMS or other computer, or each result can be sent as it is collected.

PCA-96

Guava Technologies

www.guavatechnologies.com

Compact single-cell screening

The PCA-96 AFP (Auto-Fluorescing Protein) system is a blue-laser, micro-volume cell analysis instrument that can be operated on the benchtop to screen single cells for green fluorescent protein expression using an automated 96-well plate format. The Guava system can be used to detect cells stained with both FITC-conjugated and PE-conjugated antibodies. It can be mastered with less than a day's training and requires only microvolumes of cells.

AA-6300

Shimadzu

www.shimadzu.com

It's on the beam

Shimadzu's AA-6300 atomic absorption spectrophotometer features a dynamic beam management system to ensure long-term baseline stability with low noise levels. It includes a dual detector system coupled with a



Water features: the Palintest 8000.

dual background correction function that maintains maximum sensitivity and accurate background at all wavelengths. The result is a detection capability (direct determination) of 0.1 p.p.m. Pb with flame and 0.3 p.p.b. Pb in furnace mode. The switch from flame to furnace operation can be performed quickly and without tools. Burner height is automatically set at optimal level for each element and matrix. WizAard software includes electronic signature, user administration and audit trail to ensure compliance with security regulations.

UV WinLab version 4.00

PerkinElmer

www.perkinelmer.com

Lambda series beefed up with new software

Lambda series spectrometers are now equipped with UV WinLab 4.00 software to simplify UV/Vis analysis. The software mimics the work-flow of a busy laboratory, guiding the user through simple steps for method development, results analysis and reporting functions. Built-in, password-protected access locks methods and prevents unauthorized users from changing method parameters. The UV WinLab 4.00 program archives all results and methods in a secure, encrypted relational database for easy retrieval and review. A range of intelligent querying options transforms data from a collection of individual results into useable knowledge. This speeds up the decision-making process by identifying potential problems before they occur. UV WinLab 4.00 is also available as an upgrade for the Lambda 25/35/45 and 20/40 instruments.

Ultra Evolution

Tecan

www.tecan.com

Versatile detection platform

The Ultra Evolution detection platform from Tecan measures absorbance, fluorescence and luminescence. An optional fluorescence lifetime module for samples in microplates, which may be added as an upgrade, provides

an additional solution when assay results are hindered by compound auto-fluorescence. The instrument is suitable for use by target identification or lead discovery groups, assay development teams, high-throughput screening facilities, secondary and confirmatory screening groups, or contract research/screening companies.

LALS detection guide

Viscotek

www.viscotek.com

Get the lowdown on molecular weight

Viscotek has produced a technical booklet that explores the principles of measurement and operation of the low angle light scattering (LALS) detector. Unlike multi-angle light scattering techniques that can only determine molecular weight by extrapolation, LALS measures molecular weight directly, providing direct determination of absolute molecular weights of polymers and proteins. The booklet introduces the theory of light scattering detection, explains the mathematical proof behind it and discusses the importance of the technique in scientific research, with illustrated examples of applications.

OMNIC-format libraries

Thermo Electron

www.thermo.com

Expanded range of reference material

Thermo Electron has added OMNIC-format libraries to its portfolio of spectral libraries, all of which can be searched with the company's Spectral ID search program. Spectral libraries are searchable databases used by scientists conducting materials identification and research in a range of industries, including plastics and polymers, chemicals and petrochemicals, semiconductors, food and beverage, forensics and pharmaceuticals. By matching the pattern of a specific spectrum to the reference spectra stored in the library, researchers can solve analytical problems involving quality control issues, chemical structure determination, formulation analyses and other areas. More than 50 FT-IR and Raman spectra collections are part of Thermo's portfolio, which also includes Chemical Concepts, John Wiley and NIST. The OMNIC libraries contain some of the largest collections of high-quality FT-IR and Raman reference spectra and include the Aldrich condense phase IR and Raman libraries. Most are available in high-resolution format, with spectra measured at 4 wavenumber (4 cm^{-1}) optical resolution and with data values expressed in 16-bit representation.

These notes are compiled in the Nature office from information provided by the manufacturers.



Shimadzu AA-6300: flame and furnace operation.

San Diego



Cover
Moonrise over San Diego (Michael Burke/BlackBook)

Editor, Nature:
Philip Campbell

Publisher:
Peter Collins

Supplement Editor:
Ken Howard

Senior Editor:
Natalie DeWitt

Supplement Publisher:
Sarah Greaves

Subeditor:
Colin Sullivan

Production Editor:
Maria Hodges

Art Editor:
Martin Harrison

Layouts:
Nadia Rooney

Picture Researcher:
Barbara Izdebska

Diagrams:
Charlotte Cartwright
Alisdair Macdonald

Production:
Yvonne Strong

Sponsorship:
Mark Greene

Sponsorship Layouts:
Suzanne Coleman

The Macmillan Building,
4 Crinan Street,
London N1 9XW, UK
Tel 44 (0) 20 7833 4000
e-mail: nature@nature.com
http://www.nature.com/
nature



nature publishing group

© 2003 Nature Publishing Group

In just a few decades, San Diego has changed from a sleepy military town, known mainly for its great surfing beaches, into one of the fastest-growing high-tech centres in the world. Following hot on the heels of the more mature San Francisco Bay Area and Boston regions in the United States, and giving sites in Canada, Europe and Asia a run for their money, San Diego provides an interesting lesson in how a high-tech hub takes root and expands.

In this *Nature Outlook*, we explore the factors that have shaped its success. The most important, and yet perhaps the most difficult for other regions to emulate, is San Diego's legendary 'brain trust'. The tradition of excellence at academic centres such as the University of California, San Diego, the Salk Institute for Biological Sciences and the Scripps Research Institute has provided a solid foundation for the start-up companies that have put San Diego on the map. For instance, information about cellular signalling pathways, painstakingly deduced by graduate students, postdocs and faculty members, has been used to develop drug targets. Fresh approaches suggested by physicists for beaming data through the air have led to new forms of wireless communication and enhanced the region's already strong reputation for telecommunications.

Once the seed of a new technology has been given life in the laboratory, San Diego-based venture capitalists and entrepreneurs are extremely active in aiding the next key step. They provide the financial backing necessary to see if the technology will yield a marketable application and they help new companies to commercialize promising technologies.

Although discovery and commercialization occur across the globe, most of the world's newest companies and products can trace their roots to just a few high-tech hubs. San Diego belongs to this group, and its famous networking environment has helped to build up a high level of investment in its own companies. The next challenge will be to ensure a steady stream of investment from around the world, to increase the diversity of revenue, and to keep the high-tech sector strong in this lean financial climate.

We are grateful for the support of all our sponsors, particularly Pfizer, as a result of which all content is available free online until March 2004 with free reprints of this supplement circulated to all *Nature Reviews* and most *Nature* research journal subscribers. As always, *Nature* carries sole responsibility for all editorial content.

Ken Howard



690 Good neighbours
Eric Niller

696 The view from the top
Ken Howard

700 Genesis of a high-tech hub
Chloe Veltman

706 Best of both worlds
Paul Smaglik

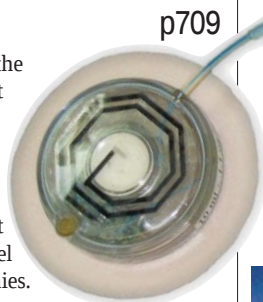
708 Turning technology into gold
Jonathan Knight

709 Tomorrow's world
Jonathan Knight

712 Different directions?
Lisa Bowman

713 In search of the élite
Virginia Gewin

720 High-tech, high society
Kendall Powell



Good neighbours

A tight-knit community and a cooperative spirit has helped San Diego to succeed. Eric Niiler checks out California's rising star.



Connected: San Diego's research network has allowed it to flourish as a high-tech hub.

Large-scale business success tends to conjure up images of a sprawling metropolis — huge areas of industrial activity linked by lengthy highways to financial hubs and disparate research institutions.

But for one rising star in the US business firmament the opposite is true. In fact, many cite San Diego's densely packed research community as a major reason for its success as a high-tech and biotech hub. Biotechnology research, for example, is located along a densely packed two-mile stretch of North Torrey Pines Road that features institutions such as the University of California, San Diego (UCSD), the Scripps Research Institute and the Salk Institute for Biomedical Studies.

"We can throw a rock and hit UCSD," says Polly Murphy, vice-president for technology management at the Salk Institute. "I can hit a golf ball and hit Scripps. Everything is within walking distance. That means more heads get together and we do a lot of collaboration."

This open attitude has helped the region to brush off the effects of economic recession in the early 1990s to become the third in terms of the number of biotech companies in the United States behind the San Francisco Bay Area and Boston (see map, overleaf). But this success has come at a price — San Diego is running out of land for development and is facing escalating house prices. As a result,

some technology firms are beginning to look elsewhere to build their new facilities. But with a high influx of federal funds, the region looks unlikely to relinquish its status just yet.

High-technology and biotechnology clusters don't grow overnight. Each follows its own long-term formula for success, usually backed by the presence of outstanding universities, a healthy flow of government research money, and venture capitalists prepared to fund start-up firms.

The cluster in San Diego is no exception, although its rapid rise has attracted the attention of scientists, entrepreneurs and policy-makers worldwide, who are keen to emulate its success in their own cities.

Strength in diversity

Today, one in eight San Diego workers is employed in the technology cluster, a broad category that includes the biotechnology, life-science and medical-device sectors, telecommunications and electronics firms, computer companies and military-related tech firms. In 2003, San Diego biotech companies reported a collective \$1.8 billion in revenue and had nearly 200 products in development, according to BIOCOM, the local biotech industry organization.

San Diego began its transformation from a tourist and military town into one of the

world's most innovative high-tech regions after a crippling recession in the early 1990s, which hit as a result of cuts in US military spending as the cold war dwindled. This led to the loss of some 60,000 jobs, but the region has managed to avoid the massive job losses in other areas during the more recent tech decline. Many other tech-dependent regions — such as the San Francisco Bay Area — continue to suffer economic headaches following the collapse of the dotcom bubble in 2001–02. Despite contractions in some fields — notably telecoms — overall employment in San Diego's tech sector has risen steadily from 99,377 in 1995 to 162,251 this year, according to Alexander X, a publisher based in San Diego that tracks such employment in California (see chart, opposite). It achieved this partly by balancing biotech with high-tech — the wireless, software and telecoms sectors have seen reduced profits over the past two years, but biotech has grown.

Another factor is that San Diego-based companies and universities have received \$1.4 billion in new defence research and development contracts since 2001 for counter-terrorism R&D. High-tech communications software and biodefence research are now priorities for the US government. Fortunately for San Diego, which suffered significant job losses in the defence sector in

and business people feel that they must share ideas and sometimes even office space.

BIOCOM has been important for promoting such informal networking. This has been particularly true for young tech entrepreneurs looking for experienced mentors who have already gone through the boom-and-bust cycles of the technology industry, says Bob Slapin, executive director of the San Diego Software Industry Council.

"The biggest difference between San Diego and other regions is a support structure and experienced individuals who are willing to assist companies," Slapin says. BIOCOM, for example, has set up joint purchasing contracts with a variety of suppliers so that a group of small biotech firms can save money by making bulk orders of specialized equipment or services such as compressed gas, lab equipment, insurance and cleaning services.

Networking and informal get-togethers have also led to a 'cross-pollination' of new technologies. For example, a blend of genomics and computational modelling has led to the evolution of several hybrid firms. Cagent Therapeutics is using three-dimensional technology originally designed to spot military targets to identify possible drug-delivery locations. "These are crossover technologies that lead from software to biology," says Fred Cutler, former head of UCSD CONNECT, an arm of the university that helps to develop and foster start-up technology firms.

Such opportunities have also proved attractive to the pharmaceutical industry. Several big drug firms have expanded their research divisions into San Diego because of

the cluster of biotech talent that exists there. Pfizer opened a new 70,000-square-metre research centre in May 2002 to study AIDS, diabetes and cancer. Novartis, Johnson & Johnson, Merck and Elan have all established West Coast research hubs in San Diego in recent years, providing more jobs in the area.

Cash rich

But if networking and cross-fertilization are driving San Diego's success, then oiling the wheels is the high level of funding available to the region. In 2002, San Diego's entire technology sector received \$1 billion in private venture-capital funding. And in a report by the Brookings Institution, the city — along with Raleigh-Durham and Seattle, two other successful biotech clusters — was identified as receiving an average of \$500 million annually from the National Institutes of Health (NIH) during 1990–2000 and at least \$700 million in private venture-capital funds from 1994–2001.

San Diego's success at attracting such high levels of funding can largely be attributed to its proactive approach. UCSD CONNECT, for example, was set up in 1995 to help start-up companies obtain financial back-up. It has subsequently initiated a series of

scientist-to-business forums that to date have helped technology companies raise more than \$5.8 billion. The various institutions in the region agree that their active solicitation of ideas from their own faculty members has also helped to bring money into the area.

"We try to make them understand that their duty is to do more than just make a

"Industry executives and other experts are warning that San Diego's technology cluster is facing some immediate challenges."

the early 1990s, it already had the infrastructure in place to meet the new challenges. "We've been saved because of our diversity," says Kelly Cunningham, an economist and research director at the San Diego Regional Chamber of Commerce.

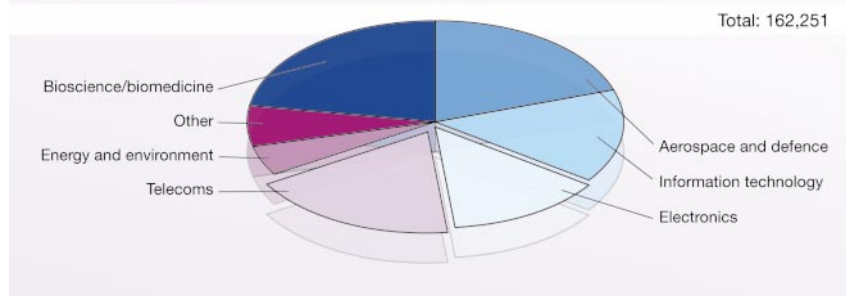
Biotech boost

Nowhere is San Diego's resurgence more obvious than in biotechnology. The city designated the 50-square-kilometre Torrey Mesa region as a zone for research and industry in the early 1960s, and since the first biotech firms were set up in the late 1980s, dozens of academic scientists have taken their ideas to nearby technology companies or started up their own.

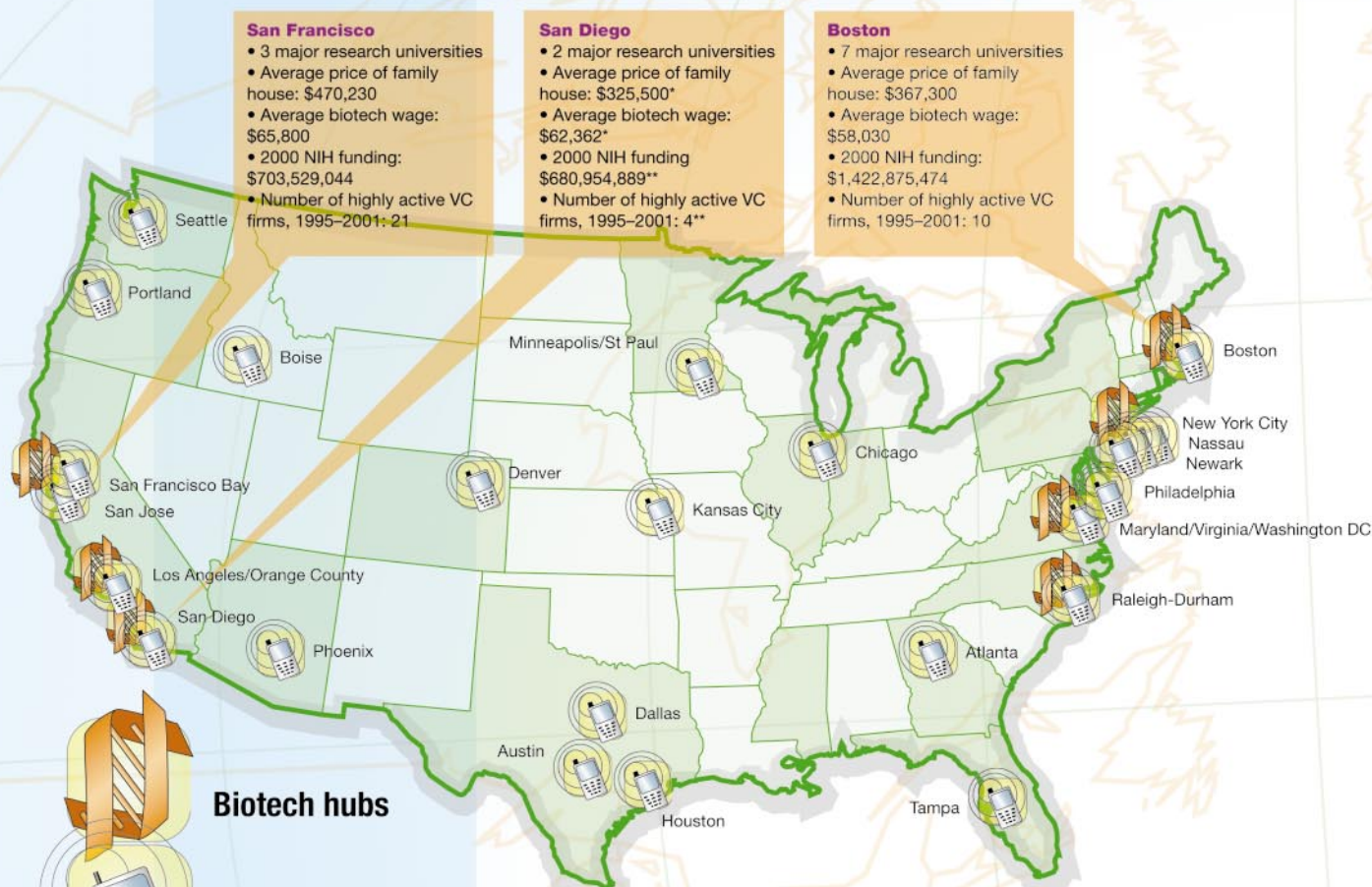
Since 1990, for example, UCSD has produced 69 biotech and high-tech firms, which currently earn the school \$17 million per year in licensing fees. The Salk Institute has launched 17 companies since the late 1980s, and researchers at Scripps — one of the largest non-profit research institutes in the United States — have developed 40 companies over a similar period of time.

Despite its rapid growth, many inhabitants maintain that San Diego still feels like a small town, living in the shadow of Los Angeles two hours' drive to the north. It is off the beaten track, which means that researchers

Who works where: the number of employees across San Diego's technology sector in 2003.



Biotech and high-tech hubs



Biotech hubs

High-tech hubs



	Value of venture-capital investments, 1995–2000 US\$ million†	Number of biotech companies, 2002*	Number of venture-capital investments, 1995–2000†	Value of pharmaceutical-biotech alliances, 1996–2001 US\$ million†	Number of initial public offerings, 1995–2000†
San Francisco Bay Area	3,029	215	261	1,205	31
Boston Area	1,916	250	219	3,924	3
San Diego	1,506	110	169	1,615	10
New York/New Jersey	639	135	63	1,729	5
Raleigh-Durham	380	85	54	192	1
Los Angeles/Orange County	181	105	26	69	1
Maryland, Virginia, Washington DC	85	110	20	358	2

Source: *Ernst & Young; †Brookings Institution

Institutions and associations in San Diego

• Institutes

Burnham Institute
California Institute for Telecommunications and Information Technology
Jacobs School of Engineering, UCSD
La Jolla Institute for Allergy and Immunology
School of Medicine, UCSD
The Neurosciences Institute
Salk Institute for Biological Studies
San Diego State University
San Diego State University Graduate School of Business
San Diego Supercomputer Center at UCSD
Scripps Clinic

Scripps Foundation for Medicine and Science
Scripps Institution of Oceanography
Sidney Kimmel Cancer Center
The Scripps Research Institute
University California, San Diego
University of San Diego

• Associations

BIOCOM
California Healthcare Institute
San Diego Regional Chamber of Commerce
San Diego Regional Economic Development Corporation
San Diego Telecom Council
UCSD CONNECT



	Number of employees, 2003 (thousands)						Total employment
	Wireless telecom	Telecom	Internet service provision	Computer and electronics manufacturing	Communications equipment manufacturing	Semiconductor manufacturing	
San Jose, California	—	—	9.3	124	—	56.3	189.6
Los Angeles/Orange County	3.4	45.2	11.1	99.6	3.4	15.1	177.8
Dallas, Texas	—	39.8	17.6	50.2	11.5	25.7	144.8
Maryland, Northern Virginia, Washington DC	3.6	59	22.3	18.9	—	—	103.8
Chicago, Illinois	—	36.8	15.5	37.7	—	—	90.0
Boston Area	—	15.4	—	48.9	5.2	11.3	80.8
San Diego	7.4	15	14.5	25.5	5.2	—	67.6
Portland, Oregon	—	—	—	34.5	—	25.3	59.8
Philadelphia, Pennsylvania	—	20.2	12.9	25.8	—	—	58.9
Atlanta, Georgia	13.4	44.6	—	—	—	—	58.0
Phoenix, Arizona	—	17.2	—	38.5	—	—	55.7
Minneapolis/St Paul, Minnesota	—	—	9.2	38.9	—	—	48.1
Austin, Texas	—	—	—	30.9	—	16.7	47.6
Seattle, Washington	10.3	18.8	—	15.7	—	—	44.8
San Francisco Bay Area	—	14.2	7.5	22.9	—	—	44.6
Houston, Texas	—	15.4	—	18.7	—	—	34.1
Tampa, Florida	—	15.8	5.3	—	3.4	3.8	28.3
New York City	—	26.2	—	—	—	—	26.2
Kansas City	—	25.8	—	—	—	—	25.8
Denver, Colorado	—	25.8	—	—	—	—	25.8
Raleigh-Durham, North Carolina	—	—	—	25.2	—	—	25.2
Boise, Idaho	—	—	—	14.5	—	9.8	24.3
Nassau, New York	—	—	—	17.8	—	4.4	22.2
Newark, New Jersey	—	13.5	—	—	—	—	13.5

Source: US Bureau of Labor Statistics, August 2003

discovery and publish a paper,” says Alan Paau, assistant vice-chancellor at UCSD and director of the school’s technology-transfer programme.

Although San Diego produces a large number of science graduates, helping to fuel its culture of innovation, relatively few of them have the business and administration experience needed to manage and expand technology firms. To help meet this need, UCSD will open a graduate business school in 2004, and has already begun a dual major PhD/MBA programme with San Diego State University. This programme allows molecular-biology students to earn a business degree at the same time as working for their main degree.

Junfu Zhang, a research fellow at the Public Policy Institute of California, says that entrepreneurs who can understand both business and science—such as those graduating from a dual programme—are more likely to succeed than those following a more traditional degree path. Zhang has studied California’s high-tech industry and found that 40% of venture-backed entrepreneurs came from a university setting. “You can never overestimate the role played by universities in developing entrepreneurship,” he says.

Saturation point

Nevertheless, industry executives and other experts are warning that San Diego’s technology cluster is facing some immediate challenges. At 50% above the national average, house prices are very high compared with wages, commuting times are increasing and the area is running out of land that can be developed. City of San Diego councillor Scott Peters admits that the region could soon become a victim of its own success. “It’s getting a little crowded,” he says.

This crowding was painfully clear during October, when wildfires spread from house to house, destroying 2,400 homes and killing 16 people in San Diego County. Fortunately for the region’s tech development, no research centre was destroyed and most analysts believe that there will not be a long-term economic effect.

Despite the damage caused by the fires in rural San Diego, the most pressing issue is the long-term provision of housing. Currently, only one in five San Diego families can afford the average house price of about \$380,000 (\$490,000 for new homes). Many young families moving to the area are forced to seek out cheaper housing in areas north and east of the city, such as Riverside county, about 60 minutes away by car. The housing

High-tech world view

Country	R&D expenditure, US\$ billion*	Number of patents filed, 1998
United States	282.3	14,401
Japan	103.8	10,230
China (excluding Hong Kong)	58.8	—
Germany	53.9	5,736
France	35.1	2,044
United Kingdom	29.4	1,851
Korea	22.3	355
India	19.4	—
Canada	17.4	511
Italy	15.5	713
Brazil	13.7	—
Russian Federation	11.6	—
Chinese Taipei	10.9	—
Sweden	9.9	951
Netherlands	8.4	782
Spain	8.2	105
Australia	7.7	271
Israel	6.4	—
Switzerland	5.6	848
Belgium	4.9	380
Finland	4.7	386
Austria	4.4	260
Mexico	3.5	12
Denmark	3.2	220

Country	R&D expenditure, US\$ billion*	Number of patents filed, 1998
Norway	2.7	117
Turkey	2.7	4
Poland	2.6	10
South Africa	2.6	—
Czech Republic	2.0	10
Singapore	2.0	—
Argentina	1.9	—
Portugal	1.5	6
Ireland	1.4	43
Hungary	1.3	24
Greece	1.1	11
Hong Kong	1.0	—
Chile	0.9	—
New Zealand	0.8	37
Slovenia	0.6	—
Romania	0.5	—
Slovak Republic	0.4	5
Iceland	0.3	11
Bulgaria	0.3	—
Lithuania	0.2	—
Estonia	0.1	—
Latvia	0.1	—
Cyprus	0.04	—

Source: OECD; *2001 or latest available year

situation is so bad that the city council declared a 'housing emergency' in 2003, and is looking for innovative ways to encourage developers to build more housing that is affordable to low- and middle-income earners.

"It's a market issue," says Peters, "and probably bigger than we can do locally. We do have to find a way to make investment in housing more attractive to private capital." One possibility, Peters says, is to require suburban developers to include in their plots a certain number of housing units that are affordable to low- and middle-income residents, although builders are distinctly uneasy about this proposal.

Another problem is the lack of water in the region. San Diego imports 80% of its water from the Colorado River 200 kilometres to the east. But the federal government has ordered California to reduce its use of the river by 20% over the next decade. San Diego officials are looking at alternatives such as desalination and recycling. Already, firms such as Ligand Pharmaceuticals, a biotech firm based in San Diego, are using reclaimed

water in their manufacturing, but the price of water is likely to add to the cost of doing business in San Diego. As a state, California is taking an increased amount of water from agricultural regions to sell to the cities. So far, it is not clear how the cost structure will pan out, although analysts agree that the price of water is going to go up.

The amount of land left to build on is shrinking, and the vaunted Torrey Mesa area is developed almost to capacity. The recent wildfires were a disquieting reminder that the wholesale real estate development in the region makes it especially vulnerable to natural disasters.

As a result, San Diego firms are starting to look for cheaper locations. IDEC Pharmaceuticals recently opened a manufacturing centre 32 kilometres north of San Diego in the city of Carlsbad (but now that it has merged with Biogen, it will move its headquarters to Massachusetts), and CancerVax is relocating its manufacturing to a Los Angeles suburb.

California officials, meanwhile, are mired in a state budget deficit of at least \$10 billion,

which could double or even triple by the middle of 2004. The state's newly elected governor, Arnold Schwarzenegger, hopes to sell bonds to plug the growing hole in the budget, but the legality of such a move is under question in the courts.

Despite the uncertainties, technology leaders in San Diego say that they are optimistic about the future. Local economic forecasters expect the region's population to grow from 2.8 million in 2000 to 3.6 million in 2020. Industry chiefs note that San Diego has grown steadily in the past 15 years by focusing on education, cross-discipline training and a cooperative business atmosphere. Although they admit that the technology cluster won't continue to grow forever, they believe that they can adapt to economic changes, natural disasters and population growth.

"We've lived with earthquakes and other disasters and people still come to California," says Joseph Panetta, executive director of BIOCOM. "We have to continue to maintain a business climate in San Diego that will allow us to continue to attract people."

Eric Niiler is a correspondent for KPBS radio in San Diego.

Profit and pitfalls: building a biotech hub

Biotechnology is now one of the prime points of focus for US regions that want to develop their local economies. In 2002 it was listed as one of the top two priorities by 83% of state and local economic-development agencies, according to a survey by the Brookings Institution. Forty-one states currently have biotech development programmes.

"Ten years ago, everyone wanted to be the next Silicon Valley; three or four years ago everyone wanted to be the centre of electronic commerce; now everyone in the economic-development community has zeroed in on biotechnology," says Joseph Cortright, an economist at Impresa, a research-analysis firm in Portland, Oregon. "That says a lot more about the herd instinct of the economic-development fraternity than it does about the economic potential and realities of biotechnology."

The problem is that biotech takes decades to develop, Cortright says, and it needs a healthy supply of venture capital to take it from research to commercialization. "If you have half-a-billion dollars, it's possible to start a research institution, but it's much more difficult, and less likely, that you will spawn a biotechnology

industry as a result," he explains.

So does a region go about creating its own biotech hub? Sceptics say that it requires not just a few big-name companies or research institutions, but a critical mass of talent and money. Bringing all of these elements together is no easy task.

Florida and South Dakota are two states that are doing their best to attract the necessary components. In late October, Florida officials approved a \$510-million package to set up a campus for the Scripps Research Institute in Palm Beach County. Based in San Diego, Scripps has an excellent record of producing spin-off companies that Florida is keen to capitalize on.

South Dakota, meanwhile, has successfully attracted biotech company Hematech, currently based in Westport, Connecticut. The company, which generates human antibodies in cattle, took advantage of the state's offer to cover half of the \$15-million costs of building a research facility in Sioux Falls.

Such moves may be helped by the rising cost of doing business in established clusters such as San Diego or the San Francisco Bay Area. "Within corporate boardrooms, there's a trend

towards smaller, more manageable, less costly locations," says John Boyd, president of corporate-relocation firm the Boyd Company, based in Princeton, New Jersey.

Boyd notes that companies also follow money. "Venture-capital firms are now being more selective about who they will fund," he says. "They would be more likely to fund a start-up in South Dakota than in San Diego because of the cost structure."

This evolution is making working conditions more flexible for prospective employees, notes John McCamant, editor of the *Medical Technology Stock Letter*, a weekly guide focusing on bioscience. "The biotech industry is big enough now and has a lot of people who can now work where they want," he says.

As an investment adviser, McCamant says that he often looks to the out-of-the-way places to find undervalued companies to invest in. He believes that biotechnology clusters could spring up just about anywhere with the help of 'tech angels', which provide unexpected assistance to start-up firms. "Most get started with angels, friends and family, and there are millionaires all over the place," he says.

E.N.

The view from the top

San Diego is facing significant challenges to its future development. How do some of the region's leading lights think it will cope? Ken Howard finds out.

How does the health of the general economy determine the fortunes of the high-tech and biotech sector? What country is the biggest challenge to the United States' dominance in these fields? Where does San Diego fit into the global high-tech enterprise? *Nature* asks some of San Diego's leading people in business and academia:

Martha Dennis, president, San Diego Telecom Council, and partner, Windward Ventures

David Gollaher, chief executive, California Healthcare Institute, La Jolla

David Hale, chief executive, CancerVax, Carlsbad

Michael Jackson, senior vice-president, the RW Johnson Pharmaceutical Research Institute, La Jolla

Tina Nova, chief executive, Genoptix

Larry Smarr, director, California Institute for Telecommunications and Information Technology, and professor, computer science and engineering, Jacobs School of Engineering, University of California, San Diego

Marco Thompson, chief technology officer, Wind River Services, San Diego and chairman, San Diego Telecom Council

How connected are the fortunes of high-tech/biotech to those of the general economy?

Martha Dennis: In general, the near-term fortunes of high-tech very closely reflect the economy. A good part of the high-tech market is determined by consumers' disposable income, and the availability of capital for new development is tied to the state of the stock market. One exception is the defence sector, which, of course, also feeds a good portion of the high-tech market, and is somewhat less vulnerable to the ups and downs of the economy.

David Gollaher: The general economy and the capital markets are the basic platform for starting new ventures, so if there's really a bear market or a market like the one we saw in 1993, there is much less risk capital available for new ventures. In a broader sense though, biotech, medical technology and medicine have the wind at their back. We have an ageing population and an increasing ability to mitigate the effects of age through medical technology, so there is an enormous demand for these products. The macroeconomy will certainly affect company formation, but the market, the ultimate demand for these products, is growing.

What is the weakest link in technology transfer out of academia?

Michael Jackson: Typically, when the first gem of an idea doesn't turn out to really go anywhere. After two years the company is no longer working on the kernel of what the founders were working on. The original idea didn't pan out, but they still have money. And they're lost.

Tina Nova: Universities should encourage development of new discoveries on campus to narrow the gap between academic science and marketable product before offering a technology or compound for licensing. Good models for this are PharmaSTART (a consortium of research organizations that offers drug-development services to California-based universities, research institutes and small biotech companies) and the von Liebig Center for Entrepreneurism and Technology Advancement (which promotes the commercialization of the work of the Jacobs School of Engineering) at the University of California, San Diego (UCSD).

Larry Smarr: Universities often have modelled their technology-transfer programmes on owning intellectual property. They should focus instead on how best to move



their ideas to market. In this model, the university's financial rewards come from the generosity of successful alumni entrepreneurs instead of revenue streams from intellectual properties. Stanford is the closest university to this ideal.

Martha Dennis: There's lack of recognition that an idea must move from discovery-focused ownership to market-focused ownership. The folk who show genius in one realm are often disasters in the other. There must be a handover from those who invent to those who build the business. The sooner universities recognize this, the faster they'll make money from their intellectual property.

How should the federal government be supporting high-tech/biotech?

Martha Dennis: Government must take on the role of filling the innovation gap that has recently been created by the shutdown or decreased funding of industry-funded research institutions such as Xerox's PARC. In the past these institutions were an essential part of maintaining the research edge that kept the United States ahead of other countries.

Tina Nova: The Small Business Administration should give Small Business Innovation Research grants to venture-backed companies. Currently, many of these grants are in jeopardy, and they can be a lifeline to struggling biotech.

Marco Thompson: Government should get out of the way. The technology and communications businesses should be deregulated as fast as possible. Government should be supporting basic research with big investments. It should not be helping the telephone service providers now in the

Left to right: Martha Dennis; David Gollaher; David Hale; Michael Jackson; Tina Nova; Larry Smarr; Marco Thompson



market. The consumers should be picking the winners.

Where does San Diego — and the United States — stand in relation to global high-tech/biotech?

David Gollaher: The United States is the global leader by far in biotech, and part of the reason is that it is the only free, non-price-controlled market in the world. Look everywhere, from the United Kingdom to Australia to Germany: all have strong price controls on drugs and medical technology so there's much less incentive in these local markets for company formation. The Europeans are waking up to see how much medical innovation they've lost, and they are very concerned.

On the world stage, you'd have to say that the Bay Area — everything from Silicon Valley up through Berkeley and down to South San Francisco — represents by far the largest cluster. The second would be the Route 128 corridor in Boston and Cambridge in Massachusetts, and the third would be San Diego. What San Diego doesn't have yet is a cluster of large, commercially successful companies that have emerging products on the market. IDEC Pharmaceuticals (developer of therapeutic monoclonal antibodies such as Rituxan) is among the first, but its products are really distributed, and to a large degree manufactured, by Genentech — and, of course, it has now combined with Biogen. So we're still looking for companies that are past the R&D stage to develop commercial manufacturing operations here in San Diego to make it a real powerhouse.

Marco Thompson: Our global leadership is threatened by the quality of our education, and the numbers of engineers graduating

elsewhere. Our university infrastructure is the best in the world, and the best foreign students still elect to attend US universities. But our high-school graduates are not even in the top ten in international surveys. Last year, US universities produced 50,000–60,000 engineering graduates. China graduated ten times that. The engineers graduating today are the entrepreneurs of tomorrow.

The main threats to our domination of high-tech come from China and India — I think principally from China. Having a great global competitor in China would be no bad thing for the United States. It will force us to take a long hard look at how we do things, and how we compete.

Martha Dennis: In certain segments of telecommunications in particular, the United States has enjoyed global technical dominance, and the telecommunications innovation companies of San Diego have been a very healthy part of this. Recently, an increased amount of design work is being done abroad. At this point, it is lower-level design work and the intellectual property that San Diego companies are developing is still critical, but there is a concern that this could change.

How does San Diego fare for money, status and talent in comparison with other high-tech hubs? Why should I start a high-tech company here?

Marco Thompson: The Bay Area is still far ahead of San Diego in all three. The venture guys are all parked on Sand Hill Road in Palo Alto and most of our entrepreneurs have to get on a one-hour flight every time they raise money. In the Bay Area, they have

been at this technology business since Hewlett and Packard set up shop in that famous garage. They have a 20-year start on the rest of the world.

San Diego will establish its status and credentials by building the fastest-growing and healthiest tech community in the world. The secret weapon is the Jacobs School of Engineering at UCSD.

Larry Smarr: San Diego is in a similar position to Silicon Valley in the late 1970s and early 1980s. There is a huge fuel tank to drive this growth in that the San Diego, Irvine and Riverside campuses of the University of California will absorb a large amount of the university's growth over the next decade. For instance, UCSD alone will hire over 400 new faculty. Companies in San Diego are also

very collaborative and organized. A good example is the San Diego Telecom Council, which has weekly talks for its nearly 300 member companies. **David Hale:** San Diego lab and office space are cheaper than in many East Coast locations, and an influx of pharmaceutical companies such as Pfizer,

Merck, Novartis and Johnson & Johnson has deepened the talent pool that all biotech companies draw from.

Martha Dennis: You'll be entering a uniquely open but close-knit technical business community known internationally for mutual support rather than back-biting. And you'll find several world-renowned educational and research institutions. You'll be able to hire staff from a rich pool of technical talent unique to San Diego. And, of course, there's always the climate. ■

Ken Howard is a contributing editor for *Nature*.

Genesis of a high-tech hub

A century of philanthropy alongside military research has laid solid foundations for today's diverse sci-tech sector, says Chloe Veltman.



Today, San Diego's North Torrey Pines Road is a congested thoroughfare lined with gleaming high-rises, low-slung modernist office blocks and towering steel cranes. It's difficult to imagine what the epicentre of the city's science and technology community must have looked like before its boom days in the early 1990s.

Inder Verma, a senior scientist at San Diego's prestigious Salk Institute for Biological Studies, has seen the landscape change beyond compare since he arrived more than two decades ago. "That used to be a hill," he says, indicating a local industrial park. "It was all wilderness. My daughter used to roam around here on her horse."

A century of academic and military research is the foundation of the city's present prosperity, both scientific and financial. It began with the first marine research centre, included the foundation of several institutes including the University of California, San Diego (UCSD), and comes up to date with the explosion of high-tech industries in the past few decades. An infrastructure for business has grown to serve sci-tech entrepreneurs, and these days even not-for-profit researchers have to think business if they want to work.

The first inklings of what was to become the science and technology hub of Torrey Pines

Mesa appeared in 1956, when the city gave defence company General Atomics (GA) 120 hectares of land on which to build a research centre. Keen to increase the high-tech presence in San Diego, city officials later designated the Mesa as a science and technology zone, allowing only high-tech organizations to build there. GA went on to spawn some 60 different science companies, including Sharp Laboratories and Science Applications International Corporation.

Long before GA arrived on the Mesa, San Diego had been priming itself to become a high-tech centre. As early as 1903, when scientist William Ritter founded the Marine Biological Association of San Diego, the city had begun to attract a research community. With funding from a newspaper mogul, E. W. Scripps, Ritter's fledgling marine laboratory transformed itself in 1912 into the Scripps Institution of Oceanography, becoming part of the University of California.

But San Diego's transformation from a town so remote that it was initially bypassed by the railroad to one of the most networked places on Earth can be traced back to the militarization of its economy during the First World War. Hoping that a military presence

would strengthen the economy, local government paid for the dredging of the harbour to accommodate navy ships. "As early as 1907, people saw an opportunity for economic growth tied to new technologies and emerging industries," says Mary Walshok, associate vice-chancellor for extended studies and public programmes at UCSD. With its non-stop sunshine and calm waters, the city soon became a favourite repository for federal defence dollars. Military facilities proliferated, bringing in a high concentration of new technologies, from long-distance radio transmission to amphibious warfare.

The military presence in San Diego proved a mixed blessing. Although healthy defence income helped lessen the blow of the Depression years and subsequent bust cycles for San Diego, and although the navy built the city's first sewage-treatment facility in 1943 and erected an aqueduct to ameliorate the area's severe water shortage, the price for the people of San Diego was high: by the Second World War, according to Mike Davis, a historian at the University of California, Irvine, San Diego had sacrificed more than a quarter of its land area to the military.

"City officials designated the Mesa a science and technology zone, allowing only high-tech organizations the privilege of building there."



Grand designs: the Neurosciences Institute (above), and the University of California's Geisel Library (top left) and geophysics lab (top) are examples of San Diego's academic environment.

If the time was ripe for the city's science to diversify, it did not have long to wait. After the war, flush with defence dollars, Scripps Institution director Roger Revelle and GA founder John Jay Hopkins wanted to establish a world-class science and engineering graduate school in the area. Revelle even offered a handful of favoured scientists, later to become UCSD's founding faculty, laboratory space at Scripps.

Teething troubles

To launch UCSD, Revelle and Hopkins had to fight a protracted battle on two fronts: with the statewide academic senate that favoured the University of California's campuses at Los Angeles and Berkeley, and with city officials who felt that UCSD should include a full undergraduate programme. UCSD finally opened in 1960 with a post-graduate science focus to support the region's new research economy. "By the time UCSD admitted undergraduates in 1964, it already had a research budget that exceeded that of a typical 100-year-old institution," says Walshok.

By then, the city's civilian scientific community had grown still larger and more diverse (see timeline, page 702). The Scripps Clinic and Research Foundation had been

founded in 1955, followed a few years later by biosciences centre the Salk Institute. Already an international hero for his Nobel prizewinning work on the polio vaccine, in 1958 Jonas Salk conceived the idea of founding an institute for sciences and humanities, a world-class research organization that would attract the top thinkers across a range of disciplines. On the lookout for a suitable site, Salk visited San Diego in 1959, and decided to build his institute there the following year. The city voted in a special referendum to donate an 11-hectare site around the corner

from the Scripps Research Institute for the project, and the National Foundation for Infantile Paralysis, which had funded Salk's polio work, pledged money.

By 1963, the first labs were set up in temporary annexes (which house faculty members to this day; see 'A modernist monastery', page 704) and the first generation of resident fellows moved in with Salk, including molecular-biology pioneers Renato Dulbecco, Francis Crick, Salvador Luria and Jacques Monod. Once this illustrious group had taken up residence on the Mesa, the future of San Diego as a scientific juggernaut was sealed.

Other key research centres followed, such as the Burnham Institute (originally called the La Jolla Cancer Research Foundation), and San Diego began to attract significant intellectual capital and funding. But it wasn't until 1968 that San Diego made the crucial connection between academia and commerce. That was when Irwin Jacobs, UCSD professor of computer science and engineering, founded Linkabit, a company that developed military signal-processing equipment and spawned Qualcomm and Leap Wireless — both still have headquarters in San Diego.

Since then, the city's academics have regularly founded companies or served as advisers to them. In 1978, for instance, Ivor Royston

Support structure: BIOCOM provides information to San Diego's life-sciences sector.

and Howard Birndorf, both UCSD scientists, founded Hybritech, the first company to commercialize the use of monoclonal antibody diagnostics. "In 1978, I succeeded in making antibodies for leukaemia, but the only way to get this to patients was to move into mass production," says Royston, who now runs local venture-capital firm Forward Ventures.

An ill-fated merger with pharmaceutical giant Eli Lilly in 1986 ultimately resulted in the sale of the then-moribund Hybritech to the medical-instruments company Beckman Coulter at a fraction of its buying price. Many of Hybritech's leading scientists left to form their own companies, such as IDEC Pharmaceuticals and Amylin: "Ultimately, what was bad for Hybritech was good for San Diego's biotech growth as a whole," says Royston.

But it was still some time before San Diego could capitalize fully on its research successes. "Business services were mismatched, focusing on defence rather than innovation," says Tony Nash, research director at New Economy Strategies, a technology-research firm in Washington DC. "San Diego lacked the intellectual-property attorneys, experienced management, and sources of capital necessary for a burgeoning economy." But with the arrival in 1985 of UCSD CONNECT, a university-based organization that helps researchers develop companies, the science community's commercial engine was shifted up a gear.

The strong relationship between business



and academia sparked capital investment, and with it grew a corps of support industries around the Torrey Pines Mesa: from real-estate brokers, specialist law firms and architects to organizations such as BIOCOM, an Internet information service for the San Diego life-sciences industry. "Companies want to be here because we have the support infrastructure here," says Brent Jacobs, senior vice-president of the life-sciences group at real-estate broker Burnham Real Estate. "They no longer have to import builders, architects and other services."

The extra momentum was soon needed. The city's dependence on the military weakened it in the early 1990s when defence cuts led to a loss of 58,000 jobs in the region, says

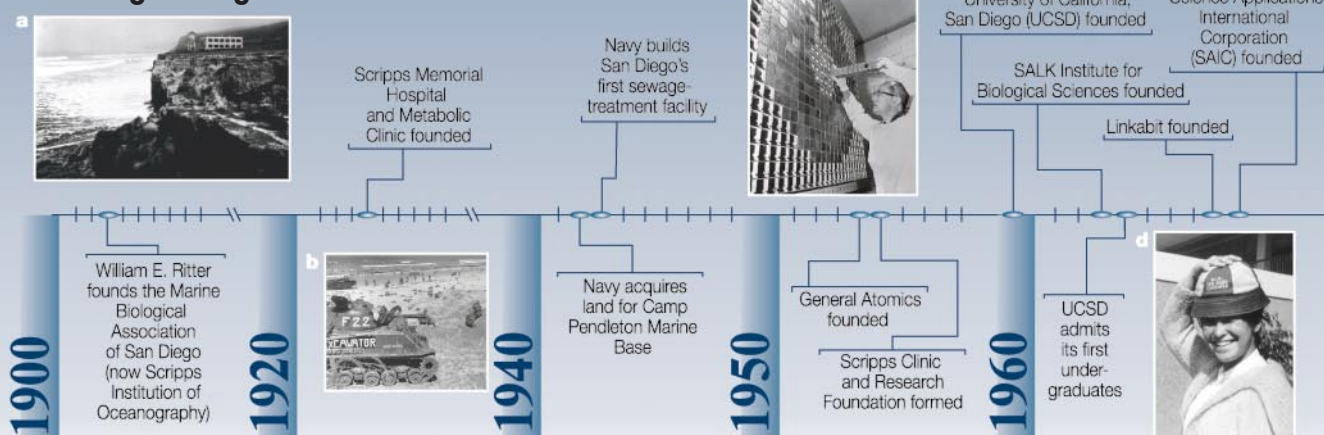
Julie Meier Wright, president and chief executive of the San Diego Regional Economic Development Corporation. By July 1993, unemployment had risen to 8.4%, whereas the national average had declined to 6.9%, according to a survey by Red Herring Research.

Variety show

Fortunately for San Diego, the fact that companies as varied as Qualcomm, General Instruments and Stratagene have made a comfortable base for themselves in San Diego owes as much to the region's roots in academic research as to its military history. The combination of decades of research and a strong entrepreneurial spirit launched the biotech, telecoms and information-technology companies that would help the region bounce back. "Since the early 1990s, our economy has diversified with industries such as wireless communications and life sciences reaching critical mass," says Wright. Through numerous boom-and-bust cycles, diversification has been key to San Diego's recovery and comparative immunity to economic downturns.

San Diego rising. a, Site of Scripps Institution of Oceanography, 1912 (Scripps Instit. Oceanogr.). b, Navy acquires site for Camp Pendleton Marine base (San Diego Hist. Soc.). c, Nuclear subcritical time-of-flight research facility at General Atomics. (CORBIS). d, UCSD's first graduate (UCSD). e, The Burnham Institute today (Burnham Inst.). f, Genetic engineer's vanity number plate (T. Spiegel/CORBIS). g, Sidney Karin, founder of UCSD's Supercomputer Center with the centre's first Cray computer (San Diego Supercomput. Center). h, Microcirculation lab at UCSD's Dept of Bioengineering (UCSD).

San Diego rising



Sun, surf...and cultural evolution

San Diego may be a major hub for technological excellence, but how does it fare in the cultural stakes? Although it dishes up a staple diet of mainstream opera, ballet and classical music, in addition to more off-the-wall arts offerings, there are moves to bring even more culture to the area.

"In the past we have been considered as culturally challenged," says Victoria Hamilton, director of the San Diego Commission for Arts and Culture, "but over the past ten years we have seen a huge increase in city investment in arts and culture." The commission's budget for 2003 was close to \$10.5 million, compared with just under \$6 million in 1993. The area now boasts a thriving arts scene ranging from museums, civic ballet, opera and orchestral groups to more eclectic offerings such as Jean Isaacs' San Diego Dance Theatre and the Sushi performance art venue. The region also has a strong San Diego-Tijuana border arts scene which includes venues such as the Centro Cultural De La Raza and the theatre company Teatro Mestizo.

Although San Diego does not yet boast many huge financial successes, riches reaped from profits and public offerings have already had a positive effect on its cultural scene. Qualcomm donates regularly to arts organizations such as the Museum of Contemporary Art San Diego and the La Jolla Playhouse. Qualcomm co-founder and chief executive Irwin Jacobs and his wife donated



On show: the windows and ceiling of the Museum of Contemporary Art in La Jolla.

\$120 million to the San Diego Symphony in 2002, a gesture that may, along with the construction of the Walt Disney Concert Hall in Los Angeles, finally catapult the southern Californian classical music scene into international prominence.

But the lure of culture comes a distant second for scientists and technology professionals

considering a move to San Diego. "Science is the main draw," says Anil Seth, a postdoctoral fellow at the Neurosciences Institute, who moved to San Diego from Britain nearly three years ago. Although the ample sunshine, outdoors lifestyle and culture certainly add to San Diego's allure, Seth adds.

C.V.

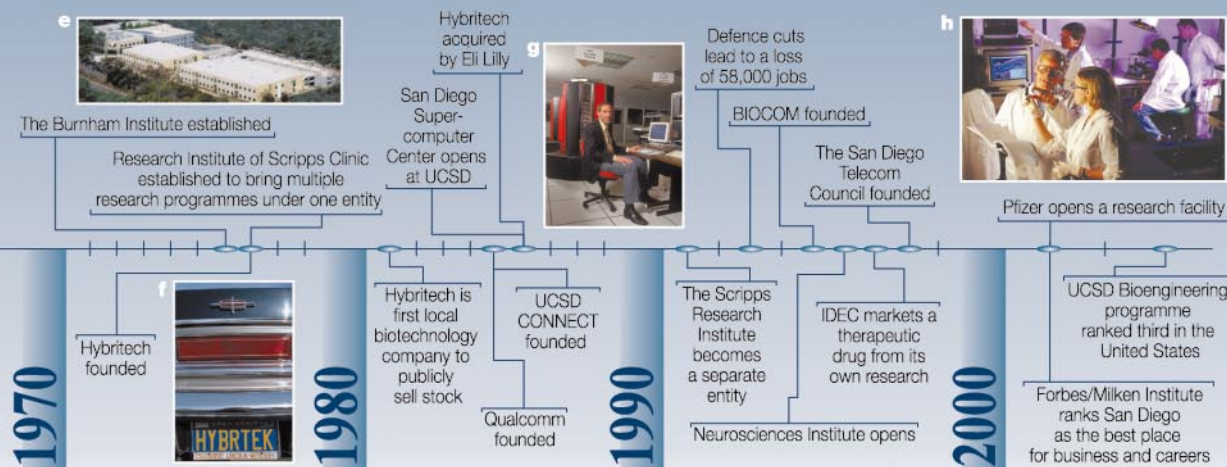
In recent years, the rate of change on the Mesa has accelerated. Drawn by the promise of breakthrough drugs and the collegial atmosphere, global pharmaceutical companies have arrived in droves. Pfizer, for example, has set up a 70,000-square-metre research facility. Others, meanwhile, have forged partnerships with smaller biotech firms as happened in Eli Lilly's collaboration with Structural GenomiX. Novartis has an agreement with Scripps in which the company pays the institute \$20 million a year for first rights to

choose up to 47% of ideas for product development until 2006. Over the past five years, major drug companies have committed more than \$1.5 billion in research and development alliances with San Diego biotech firms, according to a report from the Brookings Institution.

And a new wave of military research has come with a surge in 'biodefence' companies. Defence research has won one of the biggest increases in the 2003 federal R&D budget, accounting for nearly 55% of US research

spending, up from about 52% in 2002. There are now around 75 dedicated biodefence firms in San Diego, and other biotech companies are exploring ways to catch extra defence dollars. The biotech firm Chimerix, for example, was recently awarded a \$36-million grant from the National Institutes of Health (NIH) to develop a therapy for smallpox, classified as a 'category A' bioterrorism threat by the federal government.

But general investment in the science and technology industry and academic research



is down. According to Red Herring Research, venture-capital investment decreased by more than 30% between 2000 and 2001. "We can't fund science for its own sake," says Royston. "A few years ago, early-stage research could be funded, but financing is generally only available today for companies on their way to developing products."

The changing investment scene, and the huge California state budget deficit of at least \$10 billion, is also weighing heavily on academic organizations. "The depression within the industry has transferred the burden of translational research to academic institutions," says Ronald Evans, a long-serving scientist at the Salk Institute. "But our forte isn't commercialization, it's discovery."

With some of San Diego's academic institutions struggling to fulfil their basic research mandate in these tougher times, many researchers have begun to look for ways to bridge the funding gap. Ken Chien, director of UCSD's Institute of Molecular Medicine, believes that there will be an increasing focus on translational research — working at the interface between pure biology and its application in clinical treatments. Chien envisions increasing the number and output of physician-scientist training programmes, developing private-sector partnerships and establishing interdisciplinary teams with a range of academic and commercial skills.

San Diego's high-tech specialism has grown over many years from a unique confluence of geography, military spending, world-class academic centres and the connections fostered between the local business and research communities. "It's taken decades for San Diego to reap the benefits," says Nash. "Other regions are trying to take short-cuts but it's hard to develop excellence quickly."

San Diego faces further challenges in the coming years: increasing congestion, inadequate infrastructure — such as a lack of affordable housing — and only one small airport (see 'Good neighbours', page 690). But despite growing pains, its reputation as a high-tech haven remains high. "Everybody wants to come here," says Evans, strolling across the Salk Institute's exquisitely designed plaza towards his cosy, naturally lit office. Its previous tenant was Jonas Salk himself. "I've been here for 25 years and I don't feel much of a push to leave."

Chloe Veltman is a freelance writer in San Francisco.

A modernist monastery



The Salk Institute helped to set a high standard of architecture for San Diego's research institutes.

Torrey Pines Mesa is the most prestigious area of San Diego's high-tech hub, attracting an élite workforce with its sweeping views of the Pacific, plunging cliffs, and stunning architecture. The space is aesthetically anchored by the Salk Institute, designed by Louis Kahn, with its breezy plaza, private study towers and secluded labs. Other structures also strive to set the tone, including the sculptural Neurosciences Institute designed by New York architecture firm Tod Williams Billie Tsien & Associates in 1996. "The Salk, Scripps and the Neuroscience Institute have set a standard that all buildings seek to follow," says San Diego architect Alison Whitelaw.

Concrete, glass, water features and elegant modernist lines dominate the most notable buildings situated on the Mesa, with architects responding to the natural surroundings and climate. Inspired by the gaping canyons and spiralling vistas, architects have incorporated windows and open spaces into their designs to make the most of the views.

At the Salk Institute, a slim channel of water spans the entire length of the broad central plaza at the heart of the complex. Designed to flow from east to west, or, as Ronald Evans, a scientist at the institute, puts it, "from the known world to the unexplored frontier", the channel is directly in

line with the Sun as it traces its descent into the Pacific Ocean, drenching the plaza with late afternoon light and turning the channel into a fiery ray. Salk scientists are equally proud of the building's interior. The labs display the usual clutter of glassware and equipment, but they are

"Standing on a bluff, the old temporary annex, with its peeling linoleum and sun-bleached walls, has a ramshackle charm."

airy and open-plan, with desks arranged in a long line along one side of the room so everyone can see each other. "The architecture reflects the idea of the Mesa as a monastic setting," says Whitelaw. "Architects are engaged in creating environments that allow people to be more productive."

But for Salk scientist Joseph Ecker, the surrounding landscape is more awe-inspiring than the buildings themselves. Ecker, a plant molecular biologist, works in the old temporary annex buildings that once housed Salk's founding scientists, while the Louis Kahn building was being built. Standing on a bluff overlooking a favourite hang-gliding spot, the annex, with its peeling linoleum and sun-bleached walls, has a ramshackle charm, partly because of the vistas on its doorstep, and partly because of the sense of history and purity of purpose that permeates the concrete walls. Here, natural and artificial structures inform each other. "The juxtaposition of the two is what I think most people find so striking," says Ecker.

C.V.

Best of both worlds

What makes the Mesa so attractive to pharmaceutical companies? Paul Smaglik investigates.

The density of life-science companies and institutions in San Diego was one of the main attractions for Polly Murphy when she was looking for a new post. "If one job doesn't work out, there's a hundred right down the road," Murphy says. She also had to consider whether her husband, a psychologist, could find work in the area. But they both soon secured employment and made the move from the east coast.

Within a year, her new employer Aurora Biosciences was acquired by Vertex Pharmaceuticals in Cambridge, Massachusetts. Worried that her job might shift back to the east coast, Murphy began to look for a local alternative. True to her theory, she found a job down the road at the Salk Institute for Biological Sciences as vice-president of technology management.

In fact, Murphy could have chosen from a range of opportunities in pharmaceutical or biotechnology companies. Employment in these sectors has more than doubled between 1990 and 2000 in the San Diego area (see graph, opposite). Even in the past three years, as other US cities have shed jobs, San Diego has continued to gain them, albeit at a slower rate than during the previous decade, says Marney Cox, chief economist with regional planning agency the San Diego Association of Governments.

There are now about 400 life-science companies in the San Diego area. And even though about 90% of these firms have fewer than 100 employees, the potential for growth is already present. Many have products in phase III clinical trials, the last stage before a drug is approved to go to market. And some firms are strengthening their ties with the

area's defence industry. This could lead to an increase in biodefence funds — especially as San Diego has a good reputation for immunological research (particularly for vaccines and the detection of dangerous biological organisms such as anthrax).

But in the past decade, San Diego has also become home to some of the big names in the drug industry. Johnson & Johnson was the first big pharmaceutical firm to take an interest in the area. In the early 1980s, the company began funding research at the Scripps Research Institute on the understanding that it would have the right of first refusal if it wanted to commercialize any of the institute's discoveries. In 1994, the company recruited Per Peterson, then head of the immunology department at Scripps, to become chairman of research and development. He recruited six other scientists from Scripps in order to establish a core team in immunology.

Collaborative efforts

Johnson & Johnson's deal with Scripps ended in 1996, after which Novartis secured the offer of first refusal. Despite this, Johnson & Johnson's research in San Diego has grown from the initial seven scientists in 1994 to over 250 employees now. Its research areas include neuroscience, pain, infectious diseases and physiological systems. All teams are multidisciplinary and have biology and chemistry components.

What hasn't changed with the expansion is the close cooperation with the academic community that brought Johnson & Johnson there in the first place, says Tim Lovenberg, biology discovery team leader at the company, who describes the Torrey Pines Mesa as one giant 'open campus'.

Access to that campus for pharmaceutical companies became more important when Richard Lerner, president of Scripps, decided to change what was a "desert" for organic chemists into an oasis for them, says Mike Varney, vice-president of drug discovery at Pfizer in La Jolla. In the early 1990s, Scripps



started to build up its organic-chemistry capabilities and the University of California, San Diego, soon followed suit.

This rising strength proved attractive to Pfizer's chief executive, Hank McKinnell, when he visited the region in 2000 to see Agouron, one of the company's recent acquisitions. McKinnell was impressed with the concentration of expertise on the Mesa. So he decided to buy some land there to enhance links between Agouron's structural approach to drug design and the chemical expertise already in the area.

Agouron is now based at Pfizer's newly built San Diego facility, which is home to some 1,300 researchers. Pfizer plans to expand this workforce to about 1,500, and aims to focus on structure-based drug design at the centre using Agouron's expertise in cancer, viral diseases, diabetes and diseases of the eye.

In addition, Pfizer hopes to benefit from the surrounding academic community, which plays a supporting role in the company's research direction. Varney says that there are both "directed" and "random" components to the reasons why Pfizer and other companies pursue particular research direc-

Dual purpose: San Diego's Mesa has proved appealing to both biotech and pharmaceutical companies.

life sciences, chemistry, informatics and engineering. Employing about 450 people, the GNF splits its work equally between technology, basic biomedical research and drug discovery.

"The institute is a hybrid of academic and biotech research," says Peter Schultz, the GNF's director. This is a combination that Schultz believes gives the GNF a significant advantage. On the one hand it has technology and infrastructure not usually found at universities, on the other it has an open publication policy and can take a more opportunistic approach to research than would normally be possible at a conventional for-profit company, he says.

The GNF's hybrid nature has also allowed it to foster collaborations beyond scientists at Novartis, says Christina Waters, the institute's director of scientific development. "We can conduct independent research with other academic groups, biotech firms, and private institutions," she explains. In particular it has a strong relationship with Scripps, which is located across the road from the GNF — some GNF scientists, including Schultz, have joint faculty appointments with Scripps.

Strong connections

In fact, many companies in the region have links with academia. For example, Elan Pharmaceuticals hosts the San Diego Biotechnology Discussion Group at its Sorrento Valley location. Nabil Hanna, executive vice-president of research for Biogen Idec, says that gatherings such as the discussion group and seminars held at the academic institutions on the Mesa make recruitment easier.

But that also means opening oneself up to competition for recruits. "We hired some people from Amgen and lost some other people to Amgen," says Hanna. That sort of circulation is healthy as it allows companies to get experienced employees in management, development and regulatory affairs — experience that can't be acquired by recruiting out of universities, Hanna says.

When Stanley Crooke was looking to set up ISIS Pharmaceuticals, a company to focus on antisense therapy, he investigated several locations. Now chairman and chief executive of the company, he travelled to San Diego, liked the weather and scenery and decided to locate ISIS in Sorrento Valley, just north of La Jolla, says Shannon Devers, associate director of human resources for the firm. The ability to

attract scientists from nearby academic research institutions was a bonus, she adds.

Schering-Plough also has a small presence in San Diego through Canji, a wholly owned subsidiary. Canji has 58 employees, almost all researchers. The biotechnology company focuses on gene therapy for cancer.

But not all drug companies with a presence in San Diego are expanding. Merck, which established itself in the area when it bought SIBIA Neurosciences in August 1999, peaked at 180 employees. But as part of a companywide cutback of 4,400 employees, it is reducing that number to about 100. And last year, agribusiness giant Syngenta announced that it was closing its Torrey Mesa Research Institute in San Diego, shedding about 80 jobs, including scientists and support staff (see *Nature* **420**, 598–599; 2002).

But for every closure, there seems to be new investment. Last month, Indianapolis-based Eli Lilly announced that it would acquire Applied Molecular Evolution, a San Diego proteomics company. And earlier this year, two San Diego firms received investment from drug companies when venture capital was tight. Amylin Pharmaceuticals reached a \$400-million deal with Eli Lilly in September to market a diabetes drug currently completing the last stage of clinical trials. And Neurocrine Biosciences struck a similar deal with Pfizer last December to market the biotech firm's insomnia drug.

Murphy, meanwhile, notes that, despite all the mergers and movements of corporate headquarters from one coast to the other, she still sees a net gain for San Diego. Although she was disappointed to learn that IDEC Pharmaceuticals' corporate headquarters would shift to Massachusetts following the recent merger with Biogen, the growth of Pfizer and GNF more than make up for it.

The result of all this activity is a large variety of companies and academic institutions in a relatively contained area. That has meant more opportunities, not just for researchers but also for people providing ancillary services, such as marketing, public relations and law. Ten years ago, there were few companies specializing in these services for the life sciences. Now there are about 100.

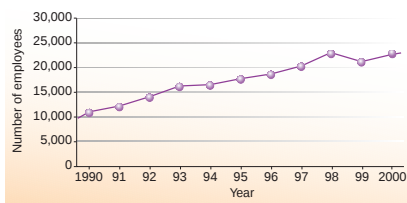
"If you're someone with skills and you want to work in San Diego, you can do it," Murphy says.

Paul Smaglik is editor of *Naturejobs*.

R. DOYLE/CORBIS

tions in the area. For example, Jerry Olefsky, a professor of medicine at the University of California, San Diego, conducted some preclinical studies for Pfizer's diabetes therapies. "This was a contributing factor to setting up diabetes and obesity as a therapeutic area at Pfizer's La Jolla site three years ago, and a useful consulting relationship now exists between the company and Olefsky," says Varney.

Pfizer and Johnson & Johnson are not alone in having close links with the academic community. In 1999, the Novartis Research Foundation established its Genomics Institute (GNF) in an effort to bring together



Rising tide: employment in San Diego's biotech and pharmaceutical sector.

Turning technology into gold

Scientists need help to make their entrepreneurial dreams come true. Jonathan Knight talks to the experts.

There are probably few scientists who have never had an entrepreneurial thought. Many researchers, at least fleetingly, have considered what it would be like to stumble across something with a real-world application and to see it move to a commercial setting.

In San Diego, as in other technology centres, discoveries flow from the laboratory bench to the marketplace with the help of specialists in technology transfer. "We are like venture capitalists, only we deal in patents," explains Alan Paau, who heads the technology-transfer office at the University of California, San Diego (UCSD).

Through licensing and other intellectual-property deals, Paau's office has created 120 start-up companies in San Diego in the past decade. These firms, combined with numerous ventures started by the other major research centres in the region, have helped to transform the town into today's high-tech hub, Paau says.

So how does technology transfer help San Diego? After all, just because a technology is invented in San Diego, it doesn't mean that it will be licensed to a San Diego firm. The top bidder for the invention could just as easily be based in Boston or Hong Kong.

Several factors keep the technology local. One is that the inventor is more readily available for consultation if companies are located close to the technology's birthplace. Another is the practice at UCSD of giving preference to local companies, even if this means forgoing a

higher bid for a licence. "As a state-run institution, we have a responsibility to focus locally," says Paau. There are also a number of organizations, such as UCSD CONNECT, keen to help start-up firms find their feet.

Before the licensing stage, someone has to decide whether the invention is worth patenting. This is where technology-transfer offices come in. Sangeeta Bhatia, a tissue engineer at UCSD who has invented a number of patented technologies, says these offices are indispensable. "Without their services, most of our biomedical innovations would be worthless," she says.

Many discoveries do not have market potential, although a researcher may find this difficult to accept. "Faculty members are in no position to judge the commercial viability of their invention," says Arnold LaGuardia, executive vice-president of the Scripps Research Institute. It is up to the licensing officer at the tech-transfer office to take a sober look at the invention and make the call.

Some research centres, such as the Salk Institute for Biological Studies, require their faculty members to report inventions to the technology-transfer office. The Scripps Research Institute goes further, reviewing every manuscript for potentially patentable discoveries before it is submitted to a journal.

More often than not, the answer from the licensing officer is no. In the fiscal year ending June 2002, the latest period for which figures are available, 255 inventions were reported to the technology-transfer office at UCSD, but only 112 patent applications were filed. One of the reasons for deciding not to patent is that a similar patent already exists. More frequently, however, it is because the licensing officer does not see any viable market for the technology.

One approach to applying for a patent is to find a company that is willing to pay the cost of filing, which usually adds up to \$25,000 for a US patent and up to ten times as much for a foreign patent. In exchange, the company gets an option to license the technology. In this model, the institution's costs are covered, and there is no reason not to patent even if a real market never materializes.

Basic research seldom leads to inventions that are obvious money-makers. Yet many early-stage innovations hold a great deal of promise for the future. "The things that will be big in the next few years were patented ten years ago," says Polly Murphy, head of technology management at the Salk Institute.

So everyone must do a certain amount of 'at-risk' filing, which means applying for a patent without a licensor. Murphy estimates

that about 40% of Salk's existing patents and patent applications are done this way, a fairly typical percentage for research institutions. John Campbell, director of intellectual property at the Burnham Institute in La Jolla agrees. "We are willing to

file a patent application and then wait until the market catches up," he says.

Usually, the patents that pay off make up for those that never find a market. For example, the Salk's patent on a research tool for inserting foreign genes into experimental organisms, filed in the mid-1990s before there was any commercial interest, has become a major revenue generator, although the institute would not say the actual amount.

Without good technology transfer, San Diego might never have seen a tech boom. As it turned out, for the scientist with an entrepreneurial bent, it's home-sweet-home. ■

Jonathan Knight is a contributing correspondent for *Nature*.

What is it worth?

Institution	Licensing revenue, 2001 (\$ million)	Per cent of all funding, 2001*	US patent applications filed, 2001	2001 patents issued	2001 licences/options executed†
Univ. California, San Diego	5.4	0.1	178	59	48
Salk Inst.	3.0	3.0	49	20	16
San Diego State Univ.	0.05	0.1	7	2	4
Scripps Inst.	N/A	N/A	N/A	N/A	N/A
Burnham Inst.	N/A	N/A	N/A	N/A	N/A
Harvard Univ.	19.1	3.7	193	37	95
Univ. California, Los Angeles	8.3	1.6	127	37	17
Univ. California, San Francisco	29.2	5.2	153	80	62
Univ. Michigan	7.9	1.2	128	65	64
Columbia Univ.	129.9	35.2	173	62	67
Univ. Washington	25.0	3.8	110	49	101

Source: Association of University Technology Managers. *Percentage of all research funding made by licensing revenue; †Number of licensing deals, including options to license, completed. N/A, not available.

Tomorrow's world

Jonathan Knight takes a look at what the future may hold for innovation in San Diego.

Predicting the future is a notoriously precarious occupation. But San Diego, with its wealth of biotech and high-tech expertise, offers a ready guide to likely successes. True, there are a lot of projects that sit firmly in the speculative category, but there are also emerging technologies that are already making their influence felt. Drawing up a comprehensive list of such developments is impossible, but from the marriage of microfabrication with cell biology to a mixture of wireless communication and medical health, a feel for the region's innovation is easy to obtain. Here, in no particular order, we present five of the hot prospects that are helping San Diego to assert itself on the global stage.

Chemical sensors

What's that smell? Is it toxic discharge from the factory up the river? Is it nerve gas from a terrorist attack? Or is it just an open rubbish bin? Increasingly, people — particularly military personnel and emergency response teams — would like to know what is in the air.

Seacoast Science based in Carlsbad is one of a handful of companies nationwide working on the next generation of accurate, lightweight chemical sensors. It has a prototype that it hopes will end up clipped to the shirts of US Army soldiers in two years' time.

The portability of the chemical detectors currently used by the US military is hampered by the fact that they are powered by batteries weighing several kilograms. In addition, the sensors are not particularly discerning. An alarm could mean a nerve-gas attack or that farmers down the road are spraying pesticides on their fields.

Part of a solution to the portability problem is to use microelectromechanical

systems: fingernail-sized devices that require only 5 microwatts for one chip of ten capacitors. Seacoast's design packs ten sensors onto a chip just 2 mm by 5 mm. Each sensor is 360 micrometres in diameter and consists of a thin layer of specialized polymer sandwiched between two conductive discs. The chemical of interest enters the sensor through perforations in the top disc and is absorbed by the polymer, altering the electrical capacitance of the polymer sandwich. This change is detected electronically.

The polymer layer is the key to this technology. Todd Mlsna, Seacoast's co-founder and president, has developed a series of absorbent polymers, each sensitive to a different class of chemical. Each of the ten sensors on a chip contains a different polymer, and any given gas changes the capacitance of the polymer sandwich at a different rate. This produces a unique electronic signature for the gas, and enables the chip to distinguish far more than ten compounds. By varying the polymers, Seacoast can tailor a chip to a given class of chemical.

In October, the company tested its sensor at the US Army Medical Research Institute of Chemical Defense in Aberdeen Proving Ground, Maryland, and showed that it could detect minute concentrations of nerve gas. Seacoast has already produced six prototypes for the Marines and is now developing ones with sensitivity to toxic industrial chemicals.

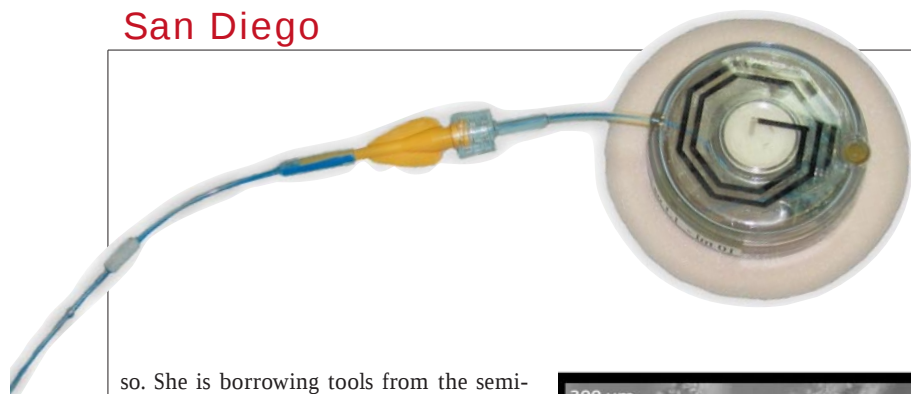
The advantage of Seacoast's sensors is their portability — every soldier can carry one. The chips do not yet distinguish nerve agents from chemically related pesticides, but this may be possible as more sensors are developed or a chemical-separation step added, according to Mlsna. "We like to think we are developing the heart of all future sensor systems," he says.

Artificial livers

Could making artificial organs one day be as simple as assembling computer chips? Sangeeta Bhatia, a bioengineer at the University of California, San Diego, thinks



Small and sensitive: Seacoast's chemical sensors can distinguish between several gases.



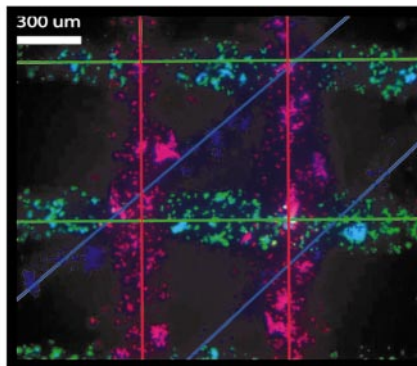
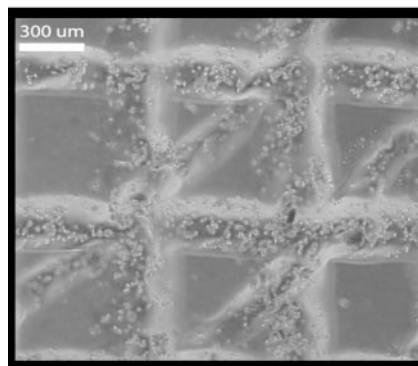
Control and construction: PhiloMetron's smart medicine patch monitors the heart and sends this information to the doctor, who can then decide how much medication to dispense (left). A three-dimensional liver-like tissue structure made from three layers of liver cells. Each layer of cells is fluorescently labelled with a different colour dye (below).

so. She is borrowing tools from the semiconductor industry to build artificial livers. So far, she has made only tiny liver fragments, but she hopes one day to build an entire lobe of liver suitable for transplanting into a patient.

Transplants are currently the only cure for patients with severe liver damage. But there aren't enough livers being donated. In the United States alone, 18,000 patients are waiting for a liver transplant, according to the American Liver Foundation, outstripping supply by fourfold.

External artificial livers, which are similar to kidney dialysis machines except that they use living cells, would extend the lives of patients waiting for a donor, but previous attempts to develop these have failed in clinical trials. One reason for this, Bhatia says, is that the liver cells tend to lose their ability to detoxify blood. To solve this problem, she grows the liver cells next to the support cells that surround them in the real organ and provide essential growth factors. She uses microfabrication tools developed for making microchips in Silicon Valley to lay down the cells in specific patterns and so control how much they interact with one another. This produces an artificial liver that is ten times more effective compared with one based on liver cells alone, she says.

But to make an entire liver suitable for transplantation, it is necessary to move away from two-dimensional cell cultures because livers are three-dimensional, not merely a layer of cells. Bhatia has begun making three-dimensional liver sections using another microfabrication method called stereolithography. This technique uses a liquid resin that is solidified layer by layer by a laser beam. For example, to make an object shaped like a tumbler, the beam sweeps back and forth across a shallow dish of resin until it forms a solid disk, which will be the bottom of the tumbler. The disk is submerged slightly deeper in the resin so that the next layer can be solidified. The laser beam must hit the liquid resin — if the disk is floating the beam will only hit the already hardened disk. Further successive rings make up the sides of tumbler. This technique can be used to make almost any three-dimensional structure.



Bhatia has developed a resin that will support growing cells (V. A. Liu and S. N. Bhatia *Biomed. Microdev.* **4**, 257–266; 2002). She mixes in liver cells, and then builds up successive layers of cells into a cube with sides of 1 mm. She leaves holes in each layer so that this small section of artificial liver has a branching vasculature like that found in real liver tissue. Bhatia uses stereolithography to make tiny liver chunks, in which the cells survive and are supported by the matrix. Work is under way to test their function.

The first application for these artificial livers could be in testing new drugs. Some 25% of new drugs fail clinical trials on people because the liver inactivates them too quickly. If this could be anticipated, the pharmaceutical industry could save a

great deal of money.

Making a functional liver for transplant is still far in the future, Bhatia says. At the moment she is using just one type of support cell, but at least three of the five major cell types found in the liver would be needed for a functional organ. This requires a much deeper understanding of how the cells interact.

Medicine patches

Only about half the patients receiving prescription medications for a chronic illness actually take them. For people with a serious condition, such as diabetes or congestive heart failure, non-compliance can mean a trip to the hospital, or worse.

Wearable devices that automatically dispense medications have been sought after for years, but very few are actually on the market.

Insulin pumps for diabetics that look like pagers and are worn on a belt are available, but they don't automatically measure blood glucose levels or communicate with the physician who is managing the disease.

PhiloMetron, a start-up company in Sorento Valley, is now taking the first steps towards tele-electronic medicine dispensers by developing what is essentially a smart skin patch that phones your doctor. The patch, which measures about 5 cm in diameter, monitors the electrical activity of the heart and sends the data at regular intervals to a nearby mobile phone, which in turn calls the doctor's office; the two steps avoid the need for a bulky transmitter in the patch itself. A clinician looks at the readings, decides how much drug to dispense, and transmits the instructions back to the skin patch, which then delivers the drug. All the patient has to do is change the patch once a week.

PhiloMetron's initial design takes advantage of an existing technology, an implanted cardio defibrillator (ICD), to gather the heart data. This is a sort of super-pacemaker that shocks the heart into a regular rhythm in the event of a heart attack. The patch can potentially go one step further too. By linking to the data provided by the ICD, it could reduce the chance of a heart attack by dispensing drugs at the appropriate time.

The next generation of patches are likely to house the data-gathering information too, so that they will work for patients who don't have an ICD fitted.

For the moment, PhiloMetron, with fewer than ten employees, remains entirely focused on the cardiac application, says president and chief executive Darrel Drinan. But it is not hard to imagine tele-medicine applications for a variety of diseases. Sensors that

"Only half the patients receiving prescriptions for chronic illness take them. Non-compliance can mean a trip to the hospital—or worse."

Light work: This unit from Maxima transmits light through air, avoiding the need for wires or cables (top). Computer models from Genomata help biologists cope with vast amounts of genomic and biochemical data (bottom).

transmit data to the clinic from wherever the patient is would not only cut down on visits to the doctor and the hospital, they would allow physicians to spot problems before they became severe.

During the severe acute respiratory syndrome (SARS) epidemic, in which the first sign of infection was fever, Drinan says he had a number of e-mails from Hong Kong asking if he had anything that could remotely monitor body temperature. "The interest in this concept is extremely high," he says.

Free-space optics

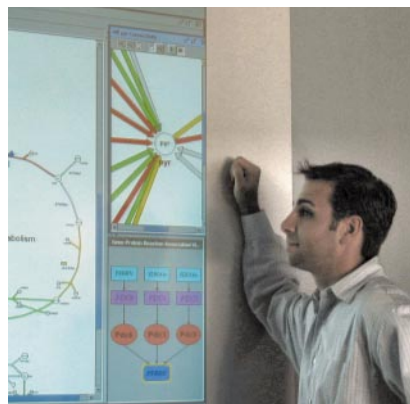
San Diego is most closely associated with one technology: wireless communication. Wireless transmission usually involves radiowaves or microwaves, but several companies in the San Diego region are developing technology based on something a little more tangible — light.

Fibre-optic cables already use light to transfer huge amounts of data. Thousands of miles of cable have been laid in the past decade, yet relatively few businesses or consumers have fibre-optic links to their desktop computers or between their private networks. This is because of the 'last mile' problem: bringing fibre-optic cables into existing buildings and houses means digging up streets and car parks. As long as existing copper wire — telephone lines and television cables — provide a reasonably fast service, the incentive to replace this technology isn't there.

Lightpointe is one of four major companies — all of which are based on the US west coast — that aim to overcome this issue by ditching the cable altogether. Free-space optics (FSO) is, in effect, fibre optics without the glass. Air is a perfectly good transmission medium for light, so sending a high-bandwidth data stream from one building to the next could be as simple as mounting a laser on one rooftop and pointing it at a receiver on another. Already there are some 10,000 FSO units installed worldwide.

Surprisingly, rain and snow do not affect the light signal. This is because Lightpointe units use four parallel laser beams to help keep the signal continuous. And lost data packets are sent again, as they usually are on the Internet. "It has to be complete white-out conditions to stop these," says Jeff Bean, spokesman for Lightpointe.

But FSO technology does have an



Achilles' heel: fog. The average size of fog droplets is one micrometre, the same as the wavelength used by standard data transmission lasers. Even a light fog can shut down FSO.

Maxima, a company based in Sorrento Valley, is trying to solve this problem by using longer wavelengths of light. Its research suggests that a wavelength of 10 micrometres would be transmitted through fog, but lasers for this wavelength are not widely available. So Maxima has joined forces with the small Swiss company Alpes Lasers, based in Neuchâtel, which has developed a laser smaller than the head of a pin that transmits a 10-micrometre beam with sufficient strength to be read hundreds of metres away. Another company, Vigo Systems in Warsaw, Poland, has designed an appropriate detector, and Maxima plans to try the

new units out on customers' rooftops by the end of 2004.

The longer wavelength has other advantages, says president and founder of Maxima James Plante. It is so far in the infrared, that sunlight has very little effect on it, so there is no interference when the Sun sets behind a data link.

In silico biology

The sequencing of a range of genomes in recent years has left many biologists with a problem. How can they make sense of so much data, particularly when the information relates to thousands of proteins involved in a slew of interdependent biochemical pathways.

Bioinformatics is helping biologists to cope with the reams of data they face every day. But few efforts are aimed at bringing all the data together into a single model to form a digital organism with the power of prediction.

Bernard Palsson, a bioengineer at the University of California, San Diego, has designed computer models of bacteria that incorporate the constraints of all their known metabolic pathways. In a paper in *Nature* last year, he and his team showed that their approach could accurately predict how a strain of the bacterium *Escherichia coli* would adapt to a new source of food (R. U. Ibarra, J. S. Edwards and B. O. Palsson *Nature* 420, 186–189; 2002).

The San Diego company Genomata, co-founded by Palsson and his former student Christophe Schilling, is now making its computerized bacteria available to researchers worldwide. The model, dubbed SimPheny, incorporates everything known about an organism's physiology, from how it metabolizes food to how it divides.

And the models aren't limited to one type of bacterium. Derek Lovley, a microbiologist at the University of Massachusetts in Amherst who works on the bacterium *Geobacter*, is a complete convert. In one case, SimPheny helped Lovley to work out which gene was mutated in a strain of *Geobacter* that was growing poorly in culture. In another, it predicted that a particular mutation would produce a strange metabolic effect that Lovley would not have tested for had the program not come up with it, he says.

Jonathan Knight is a contributing correspondent for *Nature*.

Different directions?

How independent are the boards of San Diego's high-tech companies? Lisa Bowman investigates the multiple directors.

The wave of corporate scandals that dragged down Enron also rolled over the US technology and science companies WorldCom and ImClone. Now savvy investors, corporate governance experts and potential customers are taking a close interest in who sits on a company's board of directors. Are board members independent, or do they have ties to company insiders or positions on other boards?

In the San Diego region, the board of Qualcomm, which makes wireless network equipment and is one of the region's largest tech companies, has at least four members who perform the same role elsewhere. For instance, Ray Dittamore is a director not only of Qualcomm but of the science companies Invitrogen, Gen-Probe and Applied Molecular Evolution. Other technology and science companies in the region with board-hopping directors include Anacom, SYS Technologies and Science Applications International Corporation (SAIC).

It is hard to quantify exactly what difference directors can make to research and development work in a company. A board is supposed to monitor a company, not meddle in its everyday affairs — but directors can wield significant indirect influence. Although they do not dictate what research is done, they do work with management to set strategies and allocate resources, actions that may result in one project being killed and another spawned.

Like many corporate governance experts, Charles Elson of the University of Delaware



in Newark sees no benefit to multiple directorships. "The more boards you're on, the less effective a monitor you become," he says. "It's a job that takes a lot of time." Many also question whether multiple directors are likely to challenge management decisions. "Are they getting appointed because they're diligent monitors or because they're not boat-rockers?" asks Benjamin Hermlin, a professor of banking and finance at the University of California, Berkeley.

And what do the directors get out of it? Some are corporate go-getters who enjoy helping young companies along. Others are eager for the cachet and power that accompanies a directorship. Then there's the pay: in 2002 an average board member who was not also employed by the company received about \$122,000 for the year in cash payments (retainers and per-meeting fees) and equities (grants of stock, stock options and restricted shares) according to a study of 250 companies from the Standard & Poor's 500 index, conducted by human resource consultants Towers Perrin.

But there can be some advantages to the practice too. Hermlin says that multiple directorships, although generally not recommended, could lead to collaboration between companies in the quickly changing fields of technology and science. Someone who is familiar with the industry landscape may spot opportunities or partnerships that another would overlook.

Or a director may have other valuable experience, says Larry Stambaugh, chief executive of San Diego-based Maxim Pharmaceuticals and a board member of the Corporate Directors Forum. The forum frowns upon multiple directorships, but Stambaugh says that they may make sense

in certain cases, especially if a director is an expert in accounting and oversight. For example, he points out, Dittamore retired after 35 years as a managing partner at accounting firm Ernst & Young, making him a natural target for companies seeking auditing expertise on their boards.

What's more, says Stambaugh, biotechnology start-ups — of which San Diego is a hotbed — often seek out pharmaceutical executives for their boards. "They have good access to how pharmaceutical companies are thinking, what contacts might be out there, and insight into the commercial drug process," he says. "That knowledge is helpful to a young company."

"Directors work with management to set strategies and allocate resources, actions that may result in one project being killed and another spawned."

But the general trend in the United States seems to be against multiple directorships. According to the June issue of business magazine *Forbes*, the number of people serving on five or more boards at major US companies has declined from 18 to 9 in the past year.

Many companies are also starting to limit the number of other boards their directors can sit on.

San Diego companies get particularly high marks for corporate governance. For example, 30 of the top 45 public companies in San Diego have better than average corporate governance practices, according to ratings by Institutional Shareholder Services, which examines corporate leadership for institutional investors.

Stambaugh says that directors are taking their jobs more seriously and putting in much longer hours than they did in years past. "What really changes corporate governance policies is not legislation, it's attitude, energy, putting in time and knowing how to do a good job," he emphasizes.

Lisa Bowman is a freelance writer based in San Francisco.

In search of the élite

A huge number of people have contributed to the success of San Diego. Virginia Gewin catches up with a selection of the region's prime movers.

The rise of San Diego has been remarkable. In just 30 years, it has secured itself a prime position in the world of biotech and high-tech innovation. But who helped to shape this vision? Who ensured that the region did not buckle under the pressures of recession?

There are many names on this roll of honour — those who facilitated San Diego's rejuvenation, those whose entrepreneurial vision has laid the foundations for the region's meteoric rise, and those upon whom the area's future prosperity will depend. *Nature* canvassed dozens of San Diego insiders to find out who they believed to be the élite. Among the 270 responses, several names cropped up again and again. Although by no means an exhaustive list, here, in no order of merit, are those who were mentioned most often.



Bob Beyster

Chief executive of Science Applications International Corporation (SAIC).

GREATEST ACHIEVEMENT

Founding SAIC in 1969; and 34 years of increased revenues and consistent growth for our employee owners — the company has never had a downturn year.

MOTTO Those who contribute to the company should own it.

TODAY'S TELECOM INDUSTRY It has been experiencing a downturn for the past couple years. We are optimistic that there will be a turnaround soon.

SURPRISING FACT When I started SAIC in 1969, I was never very much concerned about the money

aspect of the job. I wanted to create an environment where we would not be bothered by bureaucracy and other considerations that made it hard to do research.

BIGGEST ANNOYANCE Losing business to a competitor.

BEST PART OF JOB Solving extremely complex problems that ultimately benefit the country.



Howard Birndorf

Chief executive of Nanogen, founding partner of Hybritech and IDEC Pharmaceuticals.

SAN DIEGO AND HYBRITECH I have to believe that in time there would have been another biotech happening here because there were so many talented individuals. But it was a magical place — everybody was committed and worked hard with a sense of excellence and urgency to get the job done.

GREATEST ACCOMPLISHMENT With Hybritech it was the prostate cancer test, which has saved countless lives. I think the other great success was Rituxan, an antibody treatment that attacks cancer cells, from IDEC.

ENTREPRENEURSHIP IN SAN DIEGO It needs to be a good idea and have commercial merit, but you can get it done here easier than anywhere else in the world.

PIVOTAL MOMENT IN YOUR CAREER Meeting venture capitalist Brook Byers and working with him for 25 years. He's an amazing businessman and incredible mentor — one of the brightest people I know.

IMPACT OF NANOTECH ON BIOTECH Miniaturization is king; it will open entirely new markets.



Bob Conn

Partner at the venture-capital firm Enterprise Partners, former dean of the Jacobs School

of Engineering at the University of California, San Diego.

MOTTO You shouldn't do anything for more than a decade.

HOBBIES Contemporary art is a passionate interest. It goes back to my early days when friends were artists and I was a nerd.

PIVOTAL MOMENT OF CAREER The moments of change. I made some major changes in my career: I left the University of Wisconsin to go to the University of California, Los Angeles, then on to its San Diego campus. Finally I joined Enterprise Partners.

RECENT TRENDS As I get older I take more risks.

STATE OF HIGH-TECH IN SAN

DIEGO We're without peer in the wireless area. It is the Silicon Valley of wireless. I don't want to sound like a city booster, but I suppose I am. I think that the foundation that has been laid for San Diego's economy in the past two decades has put the city in an extraordinary position for the next ten years.



Ron Evans

Professor of molecular biology at the Salk Institute and a Howard Hughes investigator.

Founder of several companies including Ligand Pharmaceuticals, Xceptor Therapeutics and Xenopharm.

GREATEST ACHIEVEMENT The discovery of a large family of hormone receptors that are potential targets for cancer

therapy, and the characterization of a mechanism of action that revealed how hormones and drugs control wide aspects of body physiology and disease.

BIGGEST CHALLENGE TO FORMING

LIGAND Raising money. Big pharmaceutical companies weren't interested in new technology primarily because they didn't understand how it could be used.

DRIVING FORCE Helping to cure disease. Taking a discovery from the lab bench, through a company and into the clinic is amazing.

NEXT BIG THING IN BIOTECH

The next era will be the commercialization of the human genome, the integration of new concepts that will explore large numbers of proteins as potential therapeutic targets.

HOPE FOR CHANGE Some understanding by the Food and Drug Administration that every single drug should not cost \$900 million to create and test. Such a burden means that only big companies can afford to bring drugs to the market-place.

SURPRISING FACT I like fast cars.

BEST PART OF JOB Blue sky-ing with colleagues about big questions in biology.



Wain Fishburn

Partner with law firm Cooley Godward, founding member of BIOCUM and

UCSD CONNECT.

GREATEST ACHIEVEMENT Realizing in the late 1980s that the life-science and technology industries were likely to be the key factors for San Diego's future — even though at the time every one else thought I was crazy.

PIVOTAL CAREER MOMENT When San Diego matured as a life-science and technology venture community and the major law firms saw fit to open offices here. As a lawyer, I had always had to fight to demonstrate that what was happening in San Diego was important.

BEST THING ABOUT SAN DIEGO It has outgrown its caricature as a place with great sun and airheaded surfers. Over the past 20 years we have become a Mecca for the best and brightest in myriad tech fields.

SURPRISING FACT I dance when no one is looking.

TEN-YEAR FORECAST I'm curious as to whether we end up with a meaningful number of standalone companies or go in our present direction of supplying major corporations with promising companies.

FAVOURITE PLACE IN SAN DIEGO I love meeting entrepreneurs in the evening for a glass of wine on the balcony of La Valencia hotel.



Itzhak Gurantz

Chief executive of Entropic Communications, and instrumental in the development of Direct TV.

MENTOR Andy Viterbi, co-founder of Qualcomm.

GREATEST ACHIEVEMENT Developing the transmission standard and technology for Direct TV, or satellite broadcast television, especially considering its impact — and the number of doubters who questioned whether it would be valuable.

BEST PART OF JOB Working with young engineers — working and mentoring is what I enjoy.

WORST THING ABOUT SAN DIEGO Trying to recruit key executives. I wanted to hire a chief executive for a company, and a lot of those whom I considered to be qualified candidates indicated that the Bay Area was a much more favourable place for them to work.



David Hale

Chief executive of CancerVax and a founding member of Hybritech, BIOCOT and UCSD CONNECT.

HOW SAN DIEGO HAS CHANGED

When I first came here, most people had no idea what biotechnology was, what its capabilities might be or knew anything about science or industry. I remember organizing a day with the city council where a number of us presented what we thought the potential was for the industry and they were flabbergasted.

Additionally, we used to have to hire every employee from somewhere else. There wasn't anybody here that knew anything.

PERSONAL FAILURE My greatest personal disappointment was to have Protera — a product to reduce deaths and stroke during coronary bypass surgery — fail in the final stage of clinical trials at Gensia.

BIOTECH INDUSTRY FAILURE The industry as a whole has failed to grow independent biotech companies in San Diego.

FUTURE OF BIOTECH I think over the next several years, we'll have companies that can sustain growth and development in San Diego. Companies are structuring agreements so that it will make it more difficult for other companies to come and buy them.



Ed Holmes

Health sciences vice-chancellor and dean of the medical school at the University of California, San Diego.

EXCITING PROJECT PharmaSTART is the partnership between the University of California's San Diego and Los Angeles campuses, Stanford University and SRI International to move ideas or concepts out of labs and into new drugs to improve health.

ROUTE TO SUCCESS PharmaSTART

goes to the private sector with ideas to bridge the gap between discovery and the private sector. The roadblock in the 'translational highway' of moving ideas to the market-place is when an idea gets stuck at the point when its funding from the National Institutes of Health has ended, but it has yet to secure pre-seed money for a start-up. A lot of really good ideas are too early for venture capitalists to pick up and need to be matured. We want to bring an idea closer to the clinic so that the private sector might be more motivated to pick it up.

GREATEST ACHIEVEMENT Working with faculty members and leadership to develop a vision for institutions that capitalizes on their strengths and matches where the forefront of science is heading.

BEST PART OF SAN DIEGO What is unique about San Diego's academic excellence is a constellation of things — a group of institutions such as Salk, Burnham, Scripps layered on top of a robust biotech industry.

BEST PART OF JOB Bringing creativity and so making it easier for faculty members to do their research.



Irwin Jacobs

Chief executive of Qualcomm. Founded both Qualcomm and Linkabit, the company

that sparked the rise of San Diego's telecom industry and led to dozens of companies.

GREATEST ACHIEVEMENT I feel it is the development and worldwide acceptance of code division multiple access (CDMA) — digital wireless technology that creates a unique code for each call and so allows greater sharing of the airwaves.

MENTOR Claude Shannon, the father of information theory and a fellow faculty member at the Massachusetts Institute of Technology.

GREATEST BUSINESS FAILURE Our delay in entering key markets in Asia and Europe.

CURRENT STATE OF THE TELECOM INDUSTRY I think that it is slowly recovering, led by strong growth in CDMA communications, powerful and attractive devices, and a rapidly expanding range of applications.

BEST AND WORST OF SAN DIEGO There are continuing improvements in the cultural and intellectual climate to match our physical climate — although these are offset by the need for improvements in housing, transportation and public education.

SURPRISING FACT I had difficulty at college with public speaking (I almost flunked) and cooking (while in the Cornell hotel school).



Robert Kibble

Managing partner at Mission Ventures and early-stage investor in over 30 high-tech companies in California.

GREATEST BUSINESS FAILURE It's difficult to choose between them. The most recent one was the Internet company Ihome — which went bust just after the bubble burst in 2001. We just couldn't support it with additional capital.

NEXT BIG THING It's occurring right in front of our eyes — access to information from any device anywhere, complete mobility.

SAN DIEGO LACKS It would be nice to have a few more Qualcomms. And, we don't have a world-class business school. As a businessman, I see that as a negative.

HOW VENTURE CAPITAL HAS CHANGED In general, the past ten years have seen more firms and money — and it's become more institutionalized in the way it operates.

HOBBIES Until recently, horse rider and fox hunter.



Richard Lerner

President of the Scripps Research Institute.

GREATEST

ACHIEVEMENT Building Scripps into what it is today.

CHALLENGE OF RUNNING SCRIPPS

Finding the resources necessary to continue to fuel the science engine that we are today.

PIVOTAL MOMENT IN CAREER

Discovering the first catalytic antibodies in 1986.

FUTURE OF BIOTECH No matter what happens to the economy, people will always need new medicines.

NEXT BIG THING IN BIOTECH

Scripps researcher Pete Schultz's unnatural amino acids.

SURPRISING FACT I'm an excellent golfer.

NEXT FOR SCRIPPS We've just signed a deal in which we've got more than \$500 million to establish a new centre in Florida that will focus on drug discovery.

PREDICTION FOR SAN DIEGO

There's not a lot of space left, so I think the way it will evolve is that some younger companies will mature, while the research will continue to be great.

BEST PART OF YOUR JOB Getting to do my own science and interacting with the great scientists at Scripps.



Fred Muto

Partner at the law firm Cooley Godward, has worked with, or represented, over 500 technology companies in California.

CHARACTERIZE SAN DIEGO I used to live in the Bay Area and San Diego is a much more tightly knit, supportive community. The Bay Area is a federation of communities loosely held together.

CURRENT WORKLOAD I'm working with 60 tech companies right now — biotech, wireless and telecom.

CURRENT STATE OF BIOTECH The mood is very positive right now.

A lot of guarded optimism about companies and products.

FUTURE OF BIOTECH I see the future of the pharmaceutical industry as the future of the biotech industry.

CURRENT STATE OF TELECOM

Telecom remains extremely challenging. It's a huge market with a great future, but it is still a market that is continuing to contract.

BEST PART OF JOB Helping companies that make profound differences and feeling like you did a little more than practise law.



Gail Naughton

Dean of the San Diego State University Business School, founder of Advanced Tissue Sciences.

GREATEST ACCOMPLISHMENT I invented a method for growing tissue in three-dimensional scaffolding placed in a bioreactor to simulate conditions found inside the human body. This led to the formation of Advanced Tissue Sciences and to its five products for treating severe burns, diabetic wounds, genetic skin disorders and cosmetic repair.

MOST EXCITING DAY When I was the first individual woman to receive the Intellectual Property Owners Association Inventor of the Year award in 2000.

GREATEST FAILURE Not being able to keep Advanced Tissue Sciences from filing for bankruptcy in October last year. Although at the time, we had two strong marketing partnerships, three products on the market, and a fourth about to be approved, the revenues were disappointing for us and our two competitors. Within two weeks all three tissue-engineering companies with products on the market failed.

PIVOTAL MOMENT IN CAREER Deciding to form a company based on the tissue-engineering discovery even though my institution believed that the

technology could not be patented. I was told: 'If it was a good idea someone would have thought of it already' and given back rights to the technology.

SURPRISING FACT When I was a new graduate student there were two things that I knew as fact — I was never going to work in a company (never mind start one) and I was never going to have children (I now have three). I certainly learned never to say never.



Hank Nordhoff

Chief executive of Gen-Probe, board member of the cancer centre at the University of California, San Diego, and vice-chair of the Economic Regional Development Corporation.

FASTEST GROWING SECTOR FOR GEN-PROBE

Blood screening. We've developed tools to check donated blood for certain viruses including hepatitis C, and we're working on equipment to test for hepatitis B and West Nile virus.

MOST RECENT ACCOMPLISHMENT The National Institutes of Health called us last year concerned about West Nile virus and asked for a product to screen blood in time for the peak mosquito season in July 2003. We were able to turn one around and managed to intercept 400 units of blood that contained the virus.

BIGGEST FAULT I'm too slow to make personal changes.

PET ANNOYANCE People who don't use what they have as well as they could.

SURPRISING FACT I used to teach Sunday School, I've climbed Kilimanjaro and I was squash champion of Wilton, Connecticut, one year.

NEXT BIG THING IN BIOTECH Customized medicine and pharmacogenomics — being able to predict the probability of someone coming down with a disease and taking preventative measures.



Tina Nova

Chief executive and founder of GenOptix, former president and former board member

of Nanogen.

GREATEST ACHIEVEMENT Being part of teams that have successfully put products on the market that have made an impact in medical sciences. Nothing that you do can feel better than that.

PIVOTAL MOMENT OF CAREER

Deciding to go from academia to commercial industry.

HOBBIES Politics. I advise some of the political leaders on scientific matters, and I host fundraisers. I'm the California chair of the Democratic Leadership Council.

BEST ADVICE Don't let anything get in the way of what you want to accomplish.

MENTOR Somebody called me one day and said: 'I hear you are in charge of Democrats in Biotech.' I called UCSD CONNECT founder Bill Otterson and told him how strange that was, and he said: 'Oh, I made that up. We need that and I put you in charge of it.' **SURPRISING FACT** I enjoy playing *Yu-gi-oh* cards with my eight-year-old son. They're like *Pokemon* cards.

PET ANNOYANCE I prefer not to be around people who aren't passionate about something.

SAN DIEGO I wouldn't leave here for anything. You could double or triple my salary to move and I wouldn't do it.



Joe Panetta

First and current chief executive of BIOCOM.

BEST THING ABOUT

SAN DIEGO When you get out of work in the afternoon the sun's out and you can go to the beach or surf any day of the year. On the business side, I like the simple fact that people here are not only friendly but interested in your success.

WORST THING ABOUT SAN DIEGO

Although I wouldn't call it a 'worst', one of the most challenging things is the very strict environmental regulation and duplication of federal regulation in California.

BEST PART OF JOB I have the opportunity to be involved in a little bit of what's happening in 400 biotech companies at once. I spend two or three lunch hours a week visiting chief executives and touring companies — learning from them how to address their challenges. It is the greatest job in biotech in San Diego.

STATE OF THE BIOTECH INDUSTRY I really can't begin to express my enthusiasm for what biotechnology will offer in the future. Not only in terms of health, but also in food production, efficient manufacturing processes — the potential benefits are almost endless.



Bill Rastetter

Chief executive of IDEC Pharmaceuticals, who sealed the merger between Biogen and IDEC — creating the third largest biotech company in the world.

MENTOR Brook Byers of Kleiner Perkins Caufield & Byers venture-capital firm.

STATE OF BIOTECH IN SAN DIEGO Successful companies are getting bigger, and smaller companies are getting creative. The drug industry is realizing that it has to add biotech tools into its discovery or product development mix. It's a sector that has come into its own. The sector is still not profitable, but a good number of large biotech companies are and are growing rapidly.

BEST THING ABOUT SAN DIEGO Going to work and not putting on snow tyres.

SURPRISING FACT I was born in Panama.

HOBBIES Photography, swimming and golf.

GREATEST BUSINESS FAILURE The start-up technology that IDEC created did not succeed. We had to restart the company in 1993, in our seventh year, but we turned failure into success.

NEXT BIG THING IN BIOTECH

Closing of the deal between Biogen and IDEC.



John Reed

President and chief executive of the Burnham Institute, and co-founder of Idun Pharmaceuticals.

MENTOR FOR BURNHAM We have a couple of initiatives under way including an effort that has more than 100 people working on stem-cell biology. The other, the cancer drug-discovery initiative, will establish the facilities for structure-based drug discovery.

BEST THING ABOUT SAN DIEGO

The track record for public-private partnerships and entrepreneurial spirit.

PIVOTAL CAREER MOMENT While working on my MD/PhD, I went to Cold Spring Harbor Laboratory in New York for a three-week advanced cloning course and there was no more ambiguity — I wanted to do research.

MOST RECENT ACCOMPLISHMENT

I just did the Lance Armstrong Tour of Hope bike ride. Also, while finishing my PhD/MD at the University of Pennsylvania, I discovered a drug that targets a key gene known to make cancer cells immortal. The drug restores normal cell-death processes to the cancer cell and has just successfully finished the final stage of clinical trials. As a result, it seems to stand a good chance of becoming a new drug for treating cancer.

RISKY VENTURE I started Idun Pharmaceuticals based on a combination of technology from my laboratory and similar work by leaders in the field of apoptosis, or cell death. We merged the

technologies under one umbrella organization in order to get venture-capital funding. We talked to big pharmaceutical companies, but they said that it was too early to use this research for cancer treatment as it was still unproven. The only way to get ball rolling was to start a company ourselves.

SURPRISING FACT I get up at 3:00 am every morning.



Duane Roth

Chairman and chief executive of Alliance Pharmaceuticals, executive committee member for UCSD CONNECT.

CURRENT STATE OF BIOTECH IN SAN DIEGO

If you look from the skies way above and identify the hot spots of biotechnology development on the planet, San Diego is a big, bright, red light. This is the result of two factors. First, is the University of California's campus — a 40-year-old university that went from an idea to phenomenal success.

We recruited the most talented people and gave them academic freedom, which spilled over as the biotech industry started to emerge. Second, people in this area take risks — both scientific and financial.

BEST THING ABOUT SAN DIEGO We understand creative economies here. Biotech firms can have years of losses, and not make a profit until much later.

WORST THING ABOUT SAN DIEGO

The tough part for me is airplanes — it is a long way to go anywhere.

I'd like to see more financial markets move into San Diego, more capital formation here.

PIVOTAL MOMENT OF CAREER The adversity I faced on 8 January 2001, when I was told we had a potential problem that would cause us to stop working on a promising product. In the pharmaceutical industry, we live, breathe and die by our products.



Ivor Royston

Co-founder of Forward Ventures and of Hybritech.

PIVOTAL CAREER

MOMENT I was reading about monoclonal antibodies in *Nature* in 1975, and I thought: 'That's it, I'm going to use this new technology to develop new treatments for cancer'. Within six months of being at the University of California, San Diego I had made antibodies — I was the first in the area to make monoclonal antibodies.

BEST PART OF WORKING IN SAN DIEGO

There is no better place that runs the spectrum of biology. La Jolla is like one big National Institutes of Health campus. It is equal or better than the Bay, Boston or Cambridge.

GREATEST BUSINESS FAILURE I feel badly about it because friends co-invested with me. It was a company to breed severe combined immunodeficiency disease (SCID) mice. It was a disaster because of intellectual-property issues.

HISTORY I've been intimately involved with forming 15 companies. But in the early years, I was pariah in science because it was not accepted to be involved in business. Now, it is very well accepted.

UNUSUAL FACT I had a production company, Pacific West Entertainment, in Hollywood. We made some B movies, such as *SoulTaker*. It is proof that not everything I do turns to gold.



Drew Senyei

Partner in venture-capital firm Enterprise Partners.

MISCONCEPTION

ABOUT VENTURE CAPITALISM How hard it is to move technology out of the lab and into commerce. For every dollar needed for invention, you need ten for commercialization.

ORIGINAL PROFESSION I left the practice of medicine in 1987, but I've always been an entrepreneur. I started a biotech company as a third-year medical student. It was called Molecular Biosystems and was basically my 'street MBA'.

BEST THING ABOUT SAN DIEGO We have the highest per capita of Nobel laureates, most prolific publishing scientists, and a real sense of community that develops around the biotech sector.

GREATEST ACHIEVEMENT I've done 25 deals since 1987 when I joined the partnership.

ALTERNATIVE PROFESSION If I wasn't doing venture capitalism, I would like to be a rock star.

SURPRISING FACT I'm fluent in Hungarian and cook a mean chicken paprikash. Also, I'm a bit of an amateur inventor. I have 30 patents. One is the first and only test approved by the Food and Drug Administration to predict premature birth 2–3 weeks ahead of time.

PET ANNOYANCE The perception that San Francisco is the centre of the Universe.



Larry Smarr

Director of the California Institute for Telecommunications and Information

Technology (Cal-(IT)²).

GREATEST ACHIEVEMENT I think that I have consistently been a midwife for a whole series of what has become a part of the nation's information infrastructure. I argued strongly for the construction of the first national 'backbone', which connected the National Science Foundation's five supercomputer centres in 1986, and rapidly evolved, first into the NSFnet, and then into today's commercial Internet.

PRESENT AND FUTURE Technology change comes in these 10–20 year 'S' curves. I claim that what really created Silicon Valley was a co-evolution of the wired

Internet and the PC — sort of like in 1920s when it was the automobile and the road system. What is taking off now is wireless Internet and endpoints — mobile phones, sensor networks and personal digital assistants.

FUTURE FOR YOUR CAL-(IT)² We're entering a rapid deployment period for our living laboratories, which are based on campus and serve as models for society's future technology demands. We are building tools such as large-scale optical networks, wireless Internet and sensornet — real-time sensors for monitoring factors such as environmental conditions. It's unusual work for universities to undertake, and is only possible because of our close partnerships with industry.

NEXT BIG THING The partnership between industry and four California institutes — Cal-(IT)², California NanoSystems Institute, the Center for Information Technology Research in the Interest of Society, and the Institute for Bioengineering, Biotechnology and Quantitative Biomedical Research — to put in place a persistent collaborative framework across departments, campuses and between universities and industry. If this works, it will be like moving to a higher metabolic level in the evolution of life.

SURPRISING FACT I have about 15 different species of orchids in full bloom in my house at any one time, and 120 more that are resting at the orchid babysitter in town.



Bill Stensrud

Partner in venture-capital firm Enterprise Partners.

FAVOURITE PLACE IN SAN DIEGO The Torrey Pines South golf course. It is simultaneously a spectacular and profoundly humbling experience to play there.

BEST THING ABOUT SAN DIEGO A great community of innovative

and very smart engineers.

WORST THING ABOUT SAN DIEGO

The cost of living and the airport.

CURRENT STATE OF TELECOM The industry has been experiencing complete devastation for the past three years. San Diego has been somewhat protected because of its focus on wireless technologies. The future of wireless is very bright and San Diego should emerge stronger in telecoms than it was before the crash.

GREATEST BUSINESS FAILURE

I have had so many it is hard to identify a greatest. On a grand scale, I completely failed to detect — early enough — the technology bust of the past three years.

HOBBIES Music. I am past president and on the board of the San Diego Opera, and I play classical guitar — very badly, but I enjoy it.

SURPRISING FACT I was a member of a world champion tiddlywinks team.

GREATEST ACHIEVEMENT Being part of the founding management team of telecoms firm StrataCom. It was a terrific experience with a spectacular group of people. We invented the \$15-billion frame-relay industry, created tremendous shareholder value and had a lot of fun.



Marco Thompson

Chairman of the San Diego Telecom Council and chief technology officer of Wind River Systems.

GREATEST ACHIEVEMENT: Retiring from entrepreneurship in my thirties to spend time with my family. I won't start my next company until summer 2008 when my son graduates from high school.

MENTOR The late Bill Otterson, founder of UCSD CONNECT. Bill was more responsible for the economic miracle of San Diego than any other person.

CURRENT STATE OF TELECOM I call the bottom 15 May this year. By

then people were beginning to talk up revenues, to stop laying people off and some were hiring again. It has been nothing but up ever since.

HOBBIES Hockey and heleskiing — every year I take about 20 of San Diego's high-tech chief executives heleskiing in Canada.

PET ANNOYANCE We get annoyed if people do business with other companies when it could have been done locally. There are lots of services to connect people to San Diego first.

KEY TO SUCCESS It's a combination of who you know and knowing enough about people not to waste anyone's time.



Julie Meier Wright

Chief executive of the San Diego Economic Development Research Council.

GREATEST ACHIEVEMENT

California was viewed as a business-unfriendly state in the early 1990s. Under then Governor Pete Wilson's leadership we enacted wide-ranging business-climate reforms and worked diligently to help companies grow or solve problems here. The result was a stunning turnaround in California's economy.

CURRENT BUSINESS ATMOSPHERE

It's a new day in California with a new governor.

ADVICE TO THE NEW GOVERNOR

Address the budget problems responsibly so that the financial community has renewed confidence in California.

MENTOR I have been blessed with many mentors. One of the best was Pete Wilson.

WORST THING ABOUT SAN DIEGO

The most challenging thing is what exists in the biotech industry almost everywhere — inadequate capital to fund number of ideas, some of which will turn into amazing products.

SURPRISING FACT I rollerblade. ■

Virginia Gewin is a freelance writer based in Corvallis, Oregon.

High-tech, high society

Kendall Powell finds out how to make those key business connections.

At dinner parties in San Diego's high-tech neighbourhoods no one needs to ask what you do for a living — they'll have heard already on the grapevine — but everyone will want to know how you met your first venture capitalist.

For San Diego is a high-tech networker's paradise. Surrounded by the neighbourhoods of La Jolla, Torrey Pines, Del Mar and Sorrento Valley, the city's Golden Triangle is home to much of the region's high-tech industry and its support services. Just around the corner are four world-famous academic research institutions — the University of California, San Diego (UCSD); the Salk Institute for Biological Studies; the Scripps Research Institute; and the Burnham Institute. In any week, an industrious entrepreneur can easily attend two or three networking events.

And they need to. For the budding entrepreneur, making connections can win them a mentor to show them the ropes and provide crucial leads to investors. "Companies seek out people they know or have met, often through personal relationships. That's why two companies collaborate, not because they saw it in a magazine or read about it on the Internet," says Gary Pisano, a professor of technology management at Harvard Business School.

Pisano and other experts say that the bigger the local network, the easier it is for technology start-ups to find people to hire and the more likely it will be for those people to remain in the area. "If I want to do biotech, I think about Boston, San Francisco or San Diego and not a midwest city without other firms around if I lose my job," says Pisano.

And San Diego's geographical compactness and good climate give it a definite edge



when it comes to sociability. "There are only two degrees of separation in San Diego and people are very willing to connect people," says John Otterson, managing director of the San Diego office of the Silicon Valley Bank.

Small worlds

This degree of intimacy can, however, be both a blessing and a curse. "If you are 'in' in San Diego, you'll always be in — even if you have a bad idea at some point," says a former biotechnology executive, who wishes to remain anonymous. "But if you're 'out', you are out." In areas such as San Francisco and Boston, he says, more competition between venture capitalists means that it is easier for an entrepreneur to secure initial funding.

On the other hand, the cut-throat atmosphere in these other technology-driven hotspots makes cooperation between firms more difficult than in 'small town' San Diego, say locals. "When I talk with people from the San Francisco Bay Area or spend time there, they always describe that as being a more

competitive atmosphere, and San Diego as more collaborative, more open. There's a tremendous sense that we raise the boat out of the water more by working together," says Sanford Ehrlich, director of the San Diego State University Entrepreneurial Management Center.

Good connections

Events hosted by industry-specific organizations are key to getting hooked into the inner circle of the close-knit San Diego technology world. "San Diego is the best networked telecom community in the world," claims Marco Thompson, chairman of volunteer community group the San Diego Telecom Council, and chief technical officer at computer engineering and software company Wind River Services in San Diego. It is at these networking meetings, tech leaders say, that an entrepreneur with an innovation or idea for a start-up should start making the rounds, shaking hands and swapping cards.

But meetings are only the first step. After

getting your face known and picking the brains of top executives about pitching a business plan, the rookie entrepreneur must find the right venture-capital firm or investors' group. One 'unmissable' monthly event is the San Diego Venture Group's breakfast meetings, held at the Hyatt Aven-tine in La Jolla. These allow entrepreneurs to rub shoulders with executives of local venture-capital firms. Further down the line, the monthly dinner meetings of the San Diego Tech Coast Angels introduce selected small companies to a group of private investors who will individually provide seed funds of up to \$2 million.

The small size of many companies in San Diego can be a bonus. Bud Bromley, of Pasadena-based software firm ViaLogy, who has worked at molecular-detection firms in both San Francisco and San Diego, says that the latter's mix of mostly small companies means fewer layers of middle management for an entrepreneur to work his or her way through.

"You can make presentations directly to an executive in a San Diego company instead of to a third-level manager," says Bromley. Currently vice-president for business development at ViaLogy, which makes software to enhance the performance of microarrays and protein chips, Bromley networks out of his home in the Rancho Santa Fe neighbourhood (see 'Working the neighbourhood', right).

And there's no denying that San Diego's climate and geography help. San Diego workers say that they are much more likely to catch up over drinks when it doesn't involve driving across the San Francisco–Oakland Bay Bridge or struggling through a Boston snowstorm.

After hours

Although San Diego is the seventh largest city in the United States, it often feels like one big campus. This is a result of the 'ten-minute effect'. With 90% of technology companies located within ten minutes of the UCSD campus, a lot of business socializing takes place at the restaurants and hotels in the Golden Triangle and nearby.

"Tutto Mare right off Executive Drive is buzzing anytime of day, any day," Tom Jurgensen, an intellectual-property lawyer with BKF Jurgensen, offers as an example. You need to get there very early to get a table for lunch,

and you'll always see someone you know.

According to Jurgensen, the atmosphere at Tutto Mare and at Café Japengo, another Golden Triangle favourite, is younger and livelier than at traditional business restaurants in downtown La Jolla such as George's or La Valencia. "George's is where you consummate the deal or go afterward to celebrate. But the early-on talking goes on at these higher-energy places," he says. For breakfast meetings, everyone heads off to Milton's, a Jewish deli just off highway I-5 in Del Mar.

British expatriate Alistair Mann and the

group he founded, BioBrits, take over the deck at Karl Strauss Brewery Gardens, a former Japanese restaurant converted into a brewpub complete with koi pond. BioBrits, a group of about 100 Britons working in biotechnology, meets every other month for "pure socializing", says Mann, principal at Mann International, a corporate partnering consultancy. The group ranges from post-docs to chief executives and keeps them in touch not just with each other but with the UK biotech scene. These informal meetings can help junior people get to know successful senior executives, and Mann knows of at least one postdoc who has got a job through their BioBrit connections.

Surfing the network

Not surprisingly, it doesn't hurt to surf in San Diego. Otterson and his colleagues from the financial and legal communities used to hold Saturday morning business meetings at the surf breaks at Windansea or Black's Beach. "Some days we talked specifics, some days we talked concepts, and some days we just surfed," he says. The group still meets on an ad hoc basis when schedules and waves allow.

Otterson helped to found the UCSD Cancer Center's Luau & Longboard Invitational, which raises about \$200,000 a year for pilot projects in cancer research. Company teams get paired with a surfing legend, and the event concludes with a Polynesian-style luau where execs turn in their black tie for boardshorts.

On a more formal note, the Symphony at Salk event brings out Nobel laureates, academics and La Jolla society to listen to the San Diego Symphony Orchestra beneath the stars on the institute's seaside courtyard. "It's a melding of arts and sciences, fashioned after Jonas Salk's ideas, that reaches out to many different people who have interests in art, music and science," says event chair, Betty Vale.

San Diego's small-town charm, even as the city grows, remains a draw. "Merck, Novartis and Johnson & Johnson are all building millions of square feet for major research centres. They like the fact that San Diego has energy and a huge intellectual capital. But it is still a smaller community that is very interconnected," says Jurgensen. ■

Kendall Powell is a freelance science writer in Broomfield, Colorado.

Working the neighbourhood

If Bud Bromley doesn't already know you, he'll not be shy about introducing himself. Connecting with people is his way of life.

Bromley often runs into business partners and colleagues at neighbourhood events. Being a family man, those include little league and soccer games at the Rancho Santa Fe recreation fields, homeowner's association meetings and the picnic tables at the beach in Del Mar. It doesn't hurt that he lives in one of the most exclusive neighbourhoods in northern San Diego, which houses technology chief executives, serial entrepreneurs, and their lawyers and bankers.

Bromley says that you can just as easily make an important introduction on the sidelines of your kids' soccer game as at a black-tie soiree to celebrate a neighbour's fiftieth birthday.

He gained his friendly networking attitude from his years at Hewlett-Packard, he says, where he was taught "managing by walking around". The idea is to get to know your colleagues face-to-face, and just talk about what's important. Bromley thinks San Diego has broadened this philosophy to include the whole community. "My attitude is that we're all just ex-colleagues working at different companies. We're not trying to get confidential information out of them, just staying in touch with no particular agenda."

K.P.



Bud Bromley says it's good to talk.

To affinity and beyond

Antibodies are the researcher's Swiss army knife: there's a tool for every purpose, and they're made in more shapes and sizes than ever. Pete Moore and Julie Clayton report.

Antibodies not only protect us from infection, they have been exploited for years in the laboratory — in diagnostic tests, to purify proteins and as the workhorses of cell biology to detect, locate and identify cellular proteins. Designed by evolution to recognize and bind virtually any molecule that could exist, what more could we ask of them? Quite a lot, it seems. To meet the demands of today's researchers and drug developers, antibodies are being cut down to size, tweaked into new shapes, and manufactured in entirely new ways.

Antibodies as research reagents are ubiquitous. Therapeutic antibodies have had a bumpier ride, but now they're back — and here to stay. Crucially, the large pharmaceutical companies have regained interest. "Antibodies are causing a lot of excitement because there are not many new drugs out there and 30% of the biotech industry is driven by antibody technology," says Peter Hudson of CSIRO Health Sciences and Nutrition in Melbourne, Australia. Greg Winter of the MRC Laboratory of Molecular Biology in Cambridge, UK, and co-founder of antibody drug-discovery company Domantis, agrees. "A year or two back we were interested in both clinical and research applications, but it is very difficult to make money from research applications and it is clear that the funding

from investors requires us to have a pretty clear clinical focus."

Simple financial incentives drive much of the work in therapeutic antibodies. Donald Drakeman, president and chief executive of Medarex in Princeton, New Jersey, estimates that it takes some US\$20 million to get a traditional small-molecule drug ready for clinical testing, whereas this could cost only \$2 million for an antibody.

Consulting the library

The business end of an antibody molecule is the antigen-binding site. This is the part of the antibody that latches onto a specific antigenic determinant or epitope — a complementary structure on the surface of another molecule. The antigen-binding site is also known as the variable region, or variable domain, as it is the part of the antibody molecule that varies most among the billions of different antibodies that can be made, either naturally in the human body or by protein engineering. As a large biological molecule such as a protein can contain many epitopes, you can either produce monoclonal antibodies that target just one epitope on a molecule, or raise a collection of antibodies — polyclonal antibodies — that bind to a number of different epitopes on the molecule.

For therapeutic use in humans, the gold standard is a fully human monoclonal antibody. Antibodies in which the antigen-binding site is from a mouse, even when most of the rest of the antibody is human, as in chimaeric and humanized antibodies, will always risk triggering an unwanted immune response that, at best, neutralizes the therapeutic antibody and renders it ineffective. Finding the right human antibody has been made much simpler by the advent of genetic engineering and the development of vast human antibody libraries packed with potential candidates. These libraries are based on phage-display technology and store and express billions of human variable regions. Each phage carries a DNA sequence encoding a different variable region, and displays the corresponding protein on its surface, usually as part of a phage coat protein. When a potential drug target is screened against the library, only phage carrying antibodies with an affinity for the target lock on to it. The captured phage are multiplied in bacteria, and the antibody genes re-isolated, to be fashioned into genes that can be used to produce large amounts of fully functional monoclonal antibodies in mammalian cell cultures, or into smaller antibody fragments that can be produced more cheaply in yeast or bacterial culture.

PLAYING WITH THE PIECES

Regeneron of Tarryton, New York, has turned the usual way of using antibodies as therapeutics on its head. To produce therapeutic agents that will mop up unwanted cytokines or growth factors, the company has developed a hybrid molecule called a 'Trap'. This is made from the stem of an antibody fused to two receptors for the target molecule. In this

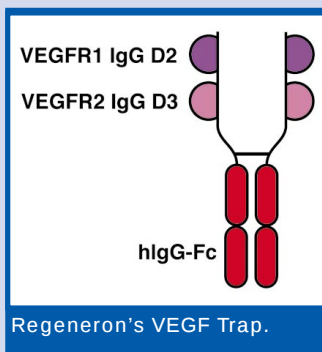
application, the antibody stem gives the molecule a long half-life in the circulation, while the receptors replace the variable regions. Traps therefore bind tightly to their target and can be given at infrequent intervals. Regeneron has created an interleukin-1 (IL-1) Trap that it hopes may be effective against rheumatoid arthritis, and is developing a dual IL-4/IL-13 Trap as a potential treatment for asthma and other allergies. A Trap for

vascular endothelial growth factor (VEGF) is in clinical trials for the treatment of cancer by preventing the growth of blood vessels into solid tumours and non-Hodgkin's lymphoma. "Every time we compare them with a conventional antibody they have affinities two or three orders of magnitude higher," says George Yancopoulos, Regeneron's chief scientific officer.

Unlike antibodies, Traps will pick up every molecule that normally binds to the receptor, so the VEGF Trap will also bind the related growth factor PLGF.

Immunotoxins — antibodies linked to cytotoxic drugs — such as Wyeth's Mylotarg, have the potential to avoid some of the toxic side effects of traditional cancer chemotherapy by delivering the drug directly to the tumour cell. Seattle Genetics of Washington state is developing SGN-15, an antibody-drug conjugate composed of a mouse-human chimaeric monoclonal antibody chemically linked to the cytotoxic drug doxorubicin. This is currently in a phase II clinical trial for the treatment of lung cancer. The antibody binds to a Lewis^x-related antigen expressed on many types of solid tumour. SGN-35, one of the company's second-generation immunotoxins now in preclinical trials, incorporates a potent synthetic drug and a novel chemical linkage that makes the molecule more stable in the blood.

P.M.



A recent success of the phage-display technology is HUMIRA, the first fully human monoclonal antibody to be approved for therapy, developed by Cambridge Antibody Technology (CAT) in Cambridge, UK, and Abbott Laboratories in Abbott Park, Illinois. This antibody, which blocks the inflammatory cytokine TNF- α , was approved in the United States in 2002 and in Europe in September this year as a treatment for the symptoms of rheumatoid arthritis.

CAT's phage-display library holds single-chain variable fragments (scFv) which consist of the variable ends of a light and a heavy chain linked with a short peptide. The library now boasts more than 100 billion distinct antigen-binding specificities. And in the world of antibody discovery, most people think that size matters. The larger the library, the greater the chance that it contains an antibody for your given target. "It means you can get reasonably high-affinity antibodies out of your first collection," says Winter.

The library started as a collection of antibody genes corresponding to antibodies

found in the blood of healthy people, but CAT has shuffled the DNA sequences that code for the variable region to give even more different types of antigen-binding site. Having started with human genes, CAT believes its antibodies are less likely to create immunogenic responses than libraries built with synthetic sequences.

Antibody discovery and production company BioInvent in Lund, Sweden, has a similar library called n-CoDeR, which holds human scFv. Like CAT, BioInvent starts with natural human antibody genes as the source of the DNA sequences, which are then further diversified by making variants of the antigen-binding regions. "We have taken evolution beyond nature and the library now has more than 10^{10} members," says Eskil Söderlind, who invented the n-CoDeR technology with Carl Borrebaeck. BioInvent can screen up to 20,000 clones per day against a customer's target antigen on its automated system. The scFv fragment genes are then converted into complete antibody genes, and the antibody itself produced in larger quantities from genetically engineered mammalian cells.

The biopharmaceutical company Dyax in Cambridge, Massachusetts, has phage-display libraries holding Fab fragments — an entire single 'arm' of an antibody. Although the company holds many of the foundation patents for phage-display technology, Dyax is a relative newcomer in applying phage display to antibodies, having initially focused on peptide and protein display. The company has moved rapidly into the antibody arena by



Going for GOLD with MorphoSys.

cross-licensing its intellectual property with other major players, such as Genentech, Xoma, Affimed, Biosite and CAT. Chief scientific officer Clive Wood believes that one of the company's strengths is the newness of its antibody libraries. "We are the new kid on the block, but because of that we have important advantages. We have been able to incorporate new features into our Fab library, such as having a mixture of natural and synthetic diversity that allows us to pull out antibodies with picomolar to nanomolar affinity." The largest Dyax library contains about 37 billion distinct human antibodies, and the company has two therapeutic proteins derived from phage display in three phase II trials.

The Human Combinatorial Antibody Library (HuCAL) of MorphoSys in Martinsried, Germany, offers 10 billion totally synthetic variable regions modelled on partial human variable regions and displayed in phage. MorphoSys has now introduced HuCAL GOLD, in which a disulphide bond

BIOINVENT



Eskil Söderlind: 10 billion antibodies.

GOING INTO PRODUCTION

"Some of the therapeutic antibodies are required now in tens of thousands of grams per year," says John Birch, chief scientific officer of the Lonza Group, based in Basel, Switzerland, which is planning three new 20,000-litre mammalian cell-culture reactors in the United States for monoclonal antibody production, in addition to its present capacity. Antibody harvests from cultured mammalian cells are around a gram per litre of culture.

But the cost of scaling-up production using mammalian cells can be prohibitive, particularly for a clinical trial of an antibody that may not turn out to be worth it. Looking to fill a gap in the market are alternatives such as the transgenic goats of GTC Biotherapeutics in Framingham, Massachusetts, and the transgenic rabbits produced under a GTC Biotherapeutics licence by BioProtein Technologies of Paris, France. "If I need it, I breed it — the walls of my bioreactor are expandable," says Thomas Newberry, vice-president of GTC Biotherapeutics. He estimates that the cost of producing a founder goat is about US\$10 million. Although pricey, this is a fraction of the estimated \$200 million–500 million that is needed to build a large-scale mammalian cell-culture bioreactor. The company estimates that a single goat can produce up to a kilogram of pure antibody per year. Goat-produced antibodies are in clinical trials to determine whether they match the safety, efficacy and pharmacokinetics of more traditionally produced counterparts.

"There's a whole category of antibodies that have not been effectively explored for therapeutic or preventive use, and for facing down viruses in the mouth, genito-urinary and respiratory tract," says Elliott Fineman, president and chief executive of Planet Biotechnology in Hayward, California. These are the secretory IgA antibodies, which cannot be produced from mammalian cell culture in significant amounts. Instead, his company is using transgenic tobacco plants to produce CaroRx, a monoclonal secretory IgA against the bacterium *Streptococcus mutans*, a leading cause of tooth decay. Originally created as 'Guys 13' by Thomas Lehner and Julian Ma at Guy's Hospital in London, CaroRx has shown its ability to block colonization by *S. mutans* in phase II clinical trials, and Planet Biotechnology is hoping to apply shortly for a licence to market the antibody in Europe for topical application by dentists and patients following antiseptic cleansing of teeth. If the company succeeds, it will be the first 'plantibody' to reach the market.

Fineman estimates that building, equipping and validating a facility to produce 100 kilograms of antibody from plants would cost between \$30 million and \$50 million. Keeping cost down is particularly important for antibodies intended for non-therapeutic use against diseases that are not life-threatening. "Nobody's going to take a £2 hike in the cost of toothpaste," says Ma.

J.C.

links the antibody to the phage instead of the antibody being an integral part of one of the coat proteins. This makes it easier to separate the antibody at a later stage. MorphoSys also offers 'Antibodies by design', a service that will develop recombinant monoclonal HuCAL antibodies for individual targets. "This is especially interesting for researchers in industry and academia, as the turnaround time is only about two months and the cost starts at about £5,000 [US\$6,000]," says Claudia Gutjahr-Loeser of MorphoSys.

Rather than going for comprehensiveness, a human-antibody library can be biased in favour of particular targets, for example by taking the initial antibody coding sequences from individuals vaccinated against a particular disease. Blood from soldiers immunized against anthrax was the starting point for a Fab antibody library that has allowed Katherine Bowdish and her

colleagues at Alexion Antibody Technologies in San Diego, California, to identify high-affinity human monoclonal antibodies to critical elements of the anthrax toxin (M. A. Wild *et al. Nature Biotechnol.* **21**, 1305–1306; 2003). The technology is also applicable to diseases such as smallpox, says Bowdish.

As the source of its library, Affitech in Oslo, Norway, also uses blood from selected donors known to produce antibodies of interest, such as cancer patients, or people who have been vaccinated or are recovering from an infectious disease. Affitech's chief executive Martin Welschof says that this, combined with the company's proprietary screening system, avoids the need for repetitive rounds of selection and means that it can move from a 10-ml blood sample to a high-affinity human monoclonal antibody of the desired specificity within four weeks.

Small is beautiful

If a therapeutic antibody has to penetrate far into body tissues, or into a tumour, for example, smaller may well be better. Although some companies are concentrating on scFv antibody fragments, UK-based Domantis is producing even smaller human antibodies from its phage libraries: these consist of just a single variable domain.

Like scFv fragments, these 'domain antibodies' can be mass-produced relatively cheaply in *Escherichia coli* or yeast. They are also very stable, which opens up new options for therapy, such as a powder that can be inhaled to treat asthma and other lung diseases, or a tablet to be taken orally for gut inflammation. "We're only going after areas where domain antibodies have unique



Frozen phage: CAT's library in store.

advantages," says Ian Tomlinson, co-founder and chief scientific officer of Domantis. In partnership with Peptech in North Ryde, New South Wales, Australia, Domantis is developing a domain antibody against TNF- α for the treatment of rheumatoid arthritis and other inflammatory diseases.

Part of the inspiration for pursuing domain antibodies was the natural domain antibodies found in camels, dromedaries, llamas — and sharks. These natural antibodies are also being explored for their potential applications by start-up biopharmaceutical company Ablynx of Zwijnaarde, Belgium.

Mouse power

Both Abgenix of Fremont, California, and Medarex of Princeton, New Jersey, take a different approach to making large numbers of different human antibodies for screening — they get mice to do it for them. In Abgenix's XenoMouse and Medarex's HuMab-Mouse, murine germ-line antibody genes have been replaced with human ones.



Katherine Bowdish: getting to grips with anthrax.

ANTIBODIES WHERE YOU WANT THEM

Antibodies are normally built inside cells, take up their final form as they are secreted from the cell, and function exclusively outside cells, binding to cell surfaces and extracellular molecules. But Terry Rabbitts at the MRC Laboratory of Molecular Biology in Cambridge, UK, wants to use antibodies' highly specific binding properties inside the cell itself. He has engineered cells to make fragments consisting of just a heavy-chain variable domain and hold them within the cell. These single-domain antibodies can interact with intracellular proteins at binding affinities of 10 nanomolar or better. At Iclectus, an MRC spin-off company based in London, Rabbitts' colleagues are developing these fragments as intracellular antibodies or 'intrabodies'.

They became interested in intrabodies while they were unravelling the role of transcription factors involved in producing chromosomal translocations in tumour cells. Using intrabodies to bind and block these transcription factors inside the cell opened up a new range of research possibilities as well as pointing the way to new anticancer therapies. Transcription factors operate by binding other proteins, so "if you have an



Iclectus: the inside story.

intrabody that has an affinity close to or better than the affinity of the interacting partner, you should be able to block the interaction using the intrabody", says Rabbitts.

Intrabodies offer the cell biologist the prospect of highly specific tools that can probe protein-protein interactions *in situ* more delicately than techniques that simply remove one of the proteins from the scene by antisense or gene knockout. Because an intrabody binds to a specific site on a protein, it can block a protein's interaction with one partner while allowing it to continue acting with another.

Another use is for the validation of potential drug targets, one of the first and vital steps in drug discovery. Most potential target proteins are modular, so intrabodies that bind to and block the function of individual modules could help to determine which part of the protein is key to causing the disease.

Placing localization signals on an intrabody allows it to be directed to various parts of the cell, such as the nucleus. "With intrabodies you can be more versatile in your experiments and ask more questions than, say, with RNA interference," says Rabbitts.

P.M.

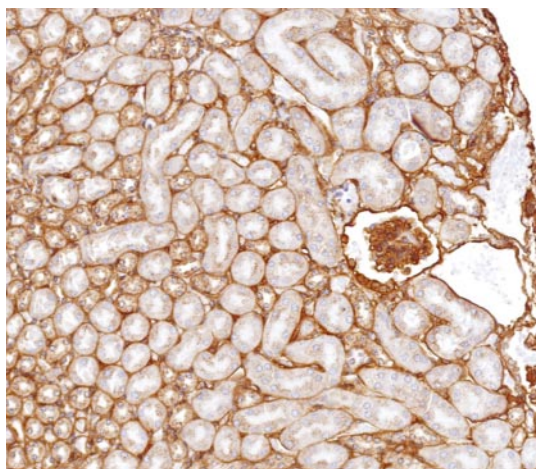
When challenged by a foreign protein, these mice produce fully human, functional antibodies. Medarex markets the HuMAb-Mouse as one part of its UltiMAb platform. “Buying the UltiMAb system allows our partners to use any of three different mice to develop products,” explains Drake-man. One is the HuMAb-Mouse. The second is Kirin TC Mouse, initially built by Kirin Brewery in Tokyo, Japan. This is a transchromosomal mouse containing the 15-megabase segment of human chromosome 14 that contains all the genes needed to build human antibodies. The third is KM-Mouse, a cross-bred version of HuMAb-Mouse and KM-Mouse. Using this technology, Medarex has generated 15 antibodies that are currently in clinical trials, while Abgenix has two in its own pipeline and others being developed by Pfizer and Amgen.

Whether you use a phage or mouse system can be influenced by what you are trying to achieve. Engineered mice don't have all the answers, claims BioInvent's Söderlind. “Mice won't react with all substances, as they may have innate tolerance to the antigen, and if the target is toxic they may die before any antibodies are formed,” he says. In those circumstances, *in vitro* phage-based systems offer a good way forward. On the other hand, say the owners of engineered mice, phage libraries can only display antibodies that do not kill bacteria.

Antibody assays

Despite a preference for human antibodies in the clinic, they are not the first choice in other settings. Monoclonal antibody technology started with mouse antibodies, and many people argue that mice are still in the driving seat. Winter recalls remarking to the late César Milstein, who created mouse monoclonal technology in the 1970s, that phage display was going to bypass the whole business of immunizing animals to get monoclonal antibodies. “He bridled at this a bit and said ‘you are wrong’ — and I think he was right,” says Winter. Milstein believed there would be a niche for all types of antibodies. And some of those niches are huge.

For when antibodies are being used as detection reagents in the laboratory, there is no need for expensive high-tech human antibodies. Rat, mouse and hamster monoclonal antibodies derived by well-established hybridoma technology, as well as polyclonal antibodies raised by immunizing larger animals, are the mainstay of laboratory immunoassays and immunodetection techniques such as ELISA (enzyme-linked immunosorbent assay) and its variant ELISpot, western blotting, flow cytometry and immunofluorescent staining of cells and tissues.



Clear view: laminin expression in kidney core.

Bender MedSystems of Vienna, Austria, offer ‘instant ELISAs’, complete ELISA kits to which the user only has to add water. These are available for a range of human, mouse and monkey immune-system cytokines and other proteins. Arrays of hundreds of antibodies are also currently the basis for the up-and-coming field of ‘protein chips’, the proteomics equivalent of DNA microarrays (see *Nature* **424**, 581–587; 2003).

Companies such as Sigma-Aldrich in St Louis, Missouri; Chemicon in Temecula, California; Active Motif in Carlsbad, California; MabTech in Nacka Strand, Sweden; and Progen Biotechnik in Heidelberg, Germany, are just some of many that supply off-the-shelf monoclonal and polyclonal antibodies against thousands of cellular targets.

Interest doesn't stop at antibodies against human and mouse proteins for biomedical research. BD Biosciences Pharmingen has thought it worthwhile to screen its large catalogue of mouse monoclonal antibodies to human intracellular structural and signalling proteins for good crossreactivity against the homologous *Drosophila* proteins. It now offers a list of about 25 high-quality monoclonal antibodies selected to give good results in *Drosophila*, and is adding more.

This wide array of laboratory reagents may soon be joined by a useful molecule that combines the binding specificity of an antibody with built-in fluorescence. These ‘fluorobodies’ would circumvent the need to add fluorescent tags to conventional antibodies. Fluorobodies consist of a super-folding version of green fluorescent protein (GFP) with the minimal antigen-binding portions of antibodies, conferring the specific protein-binding ability, grafted into four loops at one end of the GFP. They function as well as conventional antibodies in ELISAs, gel-shift assays, protein chip and flow cytometry assays — and without the need for secondary antibodies to provide a signal. “It's a once in a lifetime invention

which works really well,” says Andrew Bradbury of the Los Alamos National Laboratory in New Mexico, who is looking to commercialize the technology.

But despite the prospect of novel reagents such as fluorobodies and intrabodies, polyclonal antibodies raised in the traditional way by immunizing animals are still in demand, even for the latest protein array technologies. For example, using high-throughput techniques, Eurogentec in Seraing, Belgium, will take a customer's mRNA library and produce peptides representing different parts of the encoded proteins. Polyclonal antibodies specific for each peptide are then raised in rabbits. After selection to reduce the chance of crossreactions,

the antibodies are spotted onto glass slides to produce a customized array of up to 1,000 antibody specificities, which can be used to screen and compare tissue samples for the differential expression of multiple proteins. The approach is intended to complement the ELISA, which screens a single antibody against multiple cell or tissue samples, according to chief executive Gottfried Proess.

Antibody's-eye view

A renaissance is taking place in cellular pathology, according to Tony Warford of the Wellcome Trust's Sanger Institute in Cambridge, UK. This is thanks to the development of automated high-throughput immunohistochemistry of intact normal or diseased tissue samples. Labelled antibodies are used to stain tissues to identify the expression of particular proteins. As a senior project leader for John McCafferty's ‘Atlas of Gene Expression’ at the Sanger Institute, Warford, is pioneering the linking together of automated immunohistochemistry techniques.

By applying phage-produced monoclonal antibodies as probes to composite microarrays of tissue samples, he can see which proteins are expressed, when and where, in a given cell and tissue type. The tissue microarrays, which are created using automated technology from Beecher Instruments in Sun Prairie, Wisconsin, and Chemicon in Temecula, California, consist of paraffin wax-embedded blocks of human or animal tissues containing up to 500 individual mini-tissue samples (each about 0.6 mm across and 3 mm deep). These can then be sliced up into around 100 individual sections for immunohistochemical staining.

With automated processing, it will be possible to screen and interpret up to 20,000 individual samples — all in a day's work. ■

Pete Moore is a freelance science writer based near Bristol, UK. Julie Clayton is a science writer based in Bristol, UK.

Company	Products/activity	Location	URL
Antibody drug discovery and development			
Abbott Laboratories	Drug discovery and development; Humira therapeutic monoclonal antibody	Abbott Park, Illinois	abbott.com
Abgenix	Antibody drug discovery and development via proprietary XenoMouse and XenoMax technologies for producing fully human monoclonal antibodies	Fremont, California	www.abgenix.com
Abylnx	Development of natural single-domain antibodies (nanobodies)	Zwijnaarde, Belgium	www.ablynx.com
Agensys	Antibody drug discovery	Santa Monica, California	www.agensys.com
Affimed Therapeutics	Phage-display human antibody libraries	Heidelberg, Germany	www.affimed.com
Alexion Antibody Technologies	Antibody drug discovery and development; phage-display human Fab libraries	San Diego, California	www.alxnsd.com
Amgen	Drug discovery and development; Enbrel therapeutic monoclonal antibody	Thousand Oaks, California	www.amgen.com
Biogen idec	Drug discovery and development; therapeutic monoclonal antibodies Rituxan and Zevalin	Cambridge, Massachusetts	www.biogen.com
Bioinvent	Antibody development in proprietary n-CoDeR phage-display human scFv library	Lund, Sweden	www.bioinvent.com
Cambridge Antibody Technology	Phage-display human scFv libraries; antibody drug discovery and development	Cambridge, UK	www.cambridgeantibody.com
Centocor	Antibody and protein drug discovery and development; Remicade and ReoPro therapeutic monoclonal antibodies	Malvern, Pennsylvania	www.centocor.com
Domantis	Development of domain antibodies as therapeutics and for target validation	Cambridge, UK	www.diversys.co.uk
Dyax	Drug discovery and development; phage-display human Fab libraries	Cambridge, Massachusetts	www.dyax.com
Genentech	Antibody drug discovery and development; Xolair, Herceptin and Raptiva therapeutic monoclonal antibodies	South San Francisco, California	www.genetech.com
Iclectus	Development of intracellular antibodies (intrabodies) as therapeutics and diagnostics	London, UK	www.iclectus.com
ILEX Oncology	Drug discovery and development; CAMPATH monoclonal therapeutic antibody	San Antonio, Texas	www.ilexonc.com
ImClone	Antibody drug discovery and development; Erbitux therapeutic monoclonal antibody	New York, New York	www.imclone.com
Medarex	UltiMAB package comprising HuMAB-Mouse and other similar mice for producing fully human antibodies	Princeton, New Jersey	www.medarex.com
MorphoSys	Phage-display library of synthetic antibodies; antibody drug development	Martinsried, Germany	www.morphosys.com
Ortho Biotech	Antibody drug discovery; Orthoclone OKT3 monoclonal therapeutic antibody	Bridgewater, New Jersey	www.orthobiotech.com
Protein Design Labs	Antibody drug discovery; Zenapax monoclonal therapeutic antibody	Mountain View, California	www.pdl.com
Regeneron	Development of hybrid antibody C-region-receptor molecules (Traps) as therapeutics	Tarrytown, New York	www.regeneron.com
Xerion Pharmaceuticals	XCalibur proprietary antibody-based target-inhibition technology	Martinsried, Germany	www.xerion-pharma.com
Xoma	Antibody drug discovery and development; Raptiva monoclonal therapeutic antibody	Berkeley, California	www.xoma.com
Antibody manufacturers, suppliers and services			
AbCam	Antibody suppliers	Cambridge, UK	www.abcam.com
Abgent	Polyclonal and monoclonal antibodies; customized antibody development; off-the-shelf and custom proteins and peptides; protein-expression services	San Diego, California	www.abgent.com
Affinity BioReagents	Monoclonal and polyclonal antibodies	Golden, Colorado	www.bioreagents.com
Affitech	Phage-display human antibody libraries, screening, scFv and Fab production, regeneration to full monoclonal antibodies	Oslo, Norway	www.affitech.com
AnaSpec	Off-the-shelf and customized antibodies and peptides	San Jose, California	www.anaspec.com
Antibodies by Design	Custom monoclonal antibodies made from the proprietary Morphosys HuCAL GOLD library	Martinsried, Germany	www.antibodyservices.com
Antibodies Incorporated	Catalogue and custom antibodies; reagents and kits for immunochemistry and immunoassay	Davis, California	www.antibodiesinc.com
Aves Labs	Catalogue and custom polyclonal chicken antibodies	Tigard, Oregon	www.aveslab.com
BD Biosciences Pharmingen	Catalogue and custom antibodies; kits for immunoassays	San Diego, California	www.bdbiosciences.com/pharmingen
Biogenes	Custom monoclonal and polyclonal antibody and immunoassay development	Berlin, Germany	www.biogenes.de
Biogenesis	Antibodies and immunoassay-related reagents; bulk polyclonal and monoclonal antibodies; custom antibody development and conjugation service	Poole, UK	www.biogenesis.co.uk
Biomeda	ELISA kits, monoclonal and polyclonal antibodies	Foster City, California	biomeda.com
BioProtein Technologies	Human-antibody production in transgenic rabbits	Paris, France	www.bioprotein.com
BioSource	Custom peptides; catalogue and custom antibodies	Camarillo, California	www.biosource.com
Biotrend	Catalogue antibodies; custom antibody production service	Cologne, Germany	www.biotrend.com
Cambridge Research Biochemicals	Peptide synthesis and modification, antibody production, ELISA	Billingham, UK	www.crb.gb.com
Chemicon	Catalogue mono- and polyclonal antibodies; equipment and reagents for immunology; approved antibody-based <i>in vitro</i> diagnostic kits	Temecula, California	www.chemicon.com
Cytomix	Antibodies for kinases, receptors and ion channels	Cambridge, UK	www.cytomix.com
Delta Biolabs	Catalogue antibodies	Campbell, California	www.deltabiolabs.com
EMD Biosciences	Antibodies and other immunology tools for life-sciences research	San Diego, California	www.emdbiosciences.com
Eurogentec	Catalogue and custom antibodies; custom antibody-based protein arrays	Seraing, Belgium	www.eurogentec.be
Fusion Antibodies	Custom polyclonal, monoclonal, Fab or scFv production	Belfast, UK	www.fusionantibodies.com
Gallus Immunotech	Chicken polyclonal antibodies	Wildwood, MO	www.gallusimmunotech.com
Genetel	Antibodies for proteomics; chicken antibody specialists; antibody services	Madison, Wisconsin	www.genetel-lab.com
Genovac	Functional antibodies against G-protein-coupled receptors	Freiburg, Germany	www.genovac.com
GenWay Biotech	Primary and secondary chicken antibodies	San Diego, California	www.genwaybio.com
GTC Biotherapeutics	Human-antibody production in transgenic goats	Framingham, Massachusetts	www.gtc-bio.com
Harlan Sera-Lab	Custom monoclonal and polyclonal antibody production, peptide synthesis, purification, conjugation and hybridoma development	Loughborough, UK	www.harlanseralab.co.uk

Company	Products/activity	Location	URL
Imgenex	Catalogue and custom polyclonal and monoclonal antibodies; ELISA kits	San Diego, California	www.imgenex.com
Innovative Research	Custom and catalogue antibodies	Southfield, Michigan	www.innov-research.com
Lonza Biologics	Contract manufacture of therapeutic antibodies and recombinant proteins from mammalian cell culture	Portsmouth, New Hampshire	www.lonzabiologics.com
Mabtech	Catalogue monoclonal antibodies for cytokine detection in ELISpot, ELISA and intracellular staining; custom monoclonal production	Nacka Strand, Sweden	www.mabtech.com
Novus Biologicals	Custom and catalogue antibodies	Littleton, Colorado	www.novus-biologicals.com
Open Biosystems	Antibody services	Huntsville, Alabama	www.openbiosystems.com
PeproTech	Anti-cytokine antibodies	Rocky Hill, New Jersey	www.peprotech.com
Planet Biotechnology	Production of recombinant proteins in genetically modified plants	Hayward, California	www.planetbiotechnology.com
Progen Biotechnik	Catalogue antibodies; <i>in vitro</i> diagnostic assays	Heidelberg, Germany	www.progen.com
Research Diagnostics	Antibody distributors	Flanders, New Jersey	www.researchd.com
Serotec	Custom and catalogue antibodies	Raleigh, North Carolina	www.serotec.com
Sigma-Aldrich	Catalogue monoclonal and polyclonal antibodies; biochemicals and reagents	St Louis, Missouri	www.sigmaaldrich.com
SouthernBiotech	Catalogue and custom polyclonal and monoclonal antibodies and reagents for immunological research	Birmingham, Alabama	www.southernbiotech.com
Virusys Corporation	Catalogue and custom antibodies and antigens including chicken antibodies	Sykesville, Maryland	www.virusys.com

Immunoassay and immunochemistry

Active Motif	ELISA kits for transcription-factor assays; antibodies; peptide nucleic acids for gene silencing; kits and reagents for cell fractionation; nucleic-acid isolation	Carlsbad, California	www.activemotif.com
Alpha Diagnostics	Custom peptides and polyclonal antibodies; kits for ELISA and western blotting	San Antonio, Texas	www.4adi.com
Amersham Biosciences	Reagents for immunochemistry and immunoassays	Uppsala, Sweden	www5.amershambiosciences.com
Assay Designs	MAP kinase immunoassays (ELISA), Antibody Shop antibodies	Ann Arbor, Michigan	www.assaydesigns.com
BD Biosciences Immunocytometry Systems	FACS range of flow cytometers	San Jose, California	www.bdbiosciences.com
Beecher Instruments	Automated and manual tissue microarray instruments and accessory tissue-arrayer products	Sun Prairie, Wisconsin	www.beecherinstruments.com
Bender MedSystems	ELISA systems for research and <i>in vitro</i> diagnosis; instant ELISAs for cytokines and other immune-system molecules in humans and other species	Vienna, Austria	www.bendermedsystems.com
Bio-Rad	Products, instruments and software for life-sciences research; Bio-Plex system for multiplexing antibody-type assays; EU-approved test for BSE	Hercules, California	www.bio-rad.com
Biosite	Diagnostic antibodies; proprietary Omnidonal phage-display system for human antibody library development	San Diego, California	www.biosite.com
GenBio	ImmunoDOT and ImmunoWELL assays for clinical use; assays for antibodies characteristic of Lyme disease, autoimmunity, mononucleosis	San Diego, California	www.genbio.com
InnoGenex	Breast cancer tissue arrays; contract services for gene-expression profiling, drug-target localization and validation	San Ramon, California	www.innogenex.com
Invitrogen	Antibodies, fluorescent-tagged antibodies; immunohistochemistry kits	Carlsbad, California	www.invitrogen.com
Jerini Array Technology	Substrate identification service for kinases using pepSTAR peptide-microarray technology	Berlin, Germany	www.jerini.com
Novagen	Antibody array-based cytokine assay kit	Madison, Wisconsin	www.novagen.com
Panomics	TranSignal human cytokine antibody array	Redwood City, California	www.panomics.com
PerkinElmer Life Sciences	Automated immunoassay systems	Boston, Massachusetts	lifesciences.perkinelmer.com
Pierce Biotechnology	Components and consumables for automated systems in molecular biology, protein chemistry, sample handling, immunology and chromatography	Rockford, Illinois	www.piercenet.com
Prionics	EU-approved BSE tests	Schlieren, Switzerland	www.prionics.ch
ProSci	Catalogue and custom antibodies; immunoassays	Poway, California	www.prosci-inc.com
QED Bioscience	Catalogue and custom antibodies; immunoassay kits	San Diego, California	www.qedbio.com
R&D Systems	Cytokines; antibodies; ELISA kits; flow cytometry kits	Minneapolis, Minnesota	www.RnDSystems.com
Roche Applied Science	Reagents and kits for molecular biology, functional genomics and proteomics	Lewes, UK	www.roche-applied-science.com
Thermo Shandon	Antibodies, immunohistochemical reagents, general laboratory equipment	Runcom, UK	www.thermo.com/eThermo
Upstate	Catalogue polyclonal and monoclonal antibodies; immunoprecipitation kits	Charlottesville, Virginia	www.upstate.com
Vision Biosystems	BOND immunohistochemistry equipment and reagents	Mount Waverley, Australia	www.vision-bio.com
Zymed	Antibodies, ELISA kits, ISH kits and immunohistopathology	South San Francisco, California	www.zymed.com
Zymyx	Human-cytokine profiling biochip system	Hayward, California	www.zymyx.com

General

Calbiochem	Biochemicals and immunochemicals for life-science research	San Diego, California	www.calbiochem.com
Cambio	STAR FISH chromosome paints; suppliers of specialized biochemicals	Cambridge, UK	www.cambio.co.uk
Carl Zeiss	Imaging systems; automated systems for ultra-high throughput screening	Jena, Germany	www.zeiss.de
Chroma Technology	Filters for UV, visible and infra-red microscopy	Rockingham, Vermont	www.chroma.com
Dynal	Magnetic-bead based separation technology	Oslo, Norway	www.dynal.no
Molecular Probes	Fluorescence-based detection reagents and systems	Eugene, Oregon	www.probes.com
MP Biomedicals	Reagents and biochemicals for life-sciences research	Aurora, Ohio	www.icnbiomed.com
Peptide Specialty Labs	Custom peptides	Heidelberg, Germany	www.pslgmbh.com
USB	Biochemicals and reagents for molecular biology	Cleveland, Ohio	www.usbweb.com

● see advertisement

Contacts

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)
Andy Douglas (4975)
Frank Phelan (4944)

Netherlands/ Italy/ Spain/ Portugal/ Belgium:

Evelina Rubio Hakansson (4973)
Scandinavia: Silje Opstrup (4994)

France/ Switzerland:
Amelie Pequignot (4974)

Natureevents:
Paul Constant (4954)

Production Manager: Billie Franklin

To send materials use London
address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Office

Germany/ Austria:

Patrick Phelan, Odo Wulffen
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mail: p.phelan@nature.com
o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

US Advertising Coordinator:

Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Harakutamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami
e-mail: rasami@naturejpn.com

naturejobs

Forcing the market

A report by the National Science Foundation, issued on 24 November, paints a picture of a slowly declining US science workforce. The document, called *The Science and Engineering Workforce: Realizing America's Potential*, points to a steady slide in student enrolment for maths and physics over the past few years, projects mass retirements in key science and engineering areas over the next two decades, and predicts a decline in the number of foreign scientists who are willing — or allowed — to pick up the slack.

After reading this report, someone with no conception of the realities of the PhD job market might react with dismay, and agree with some of the conclusions that the report draws. The US education system is somehow failing. There won't be enough skilled scientists to meet the demand for qualified positions. The US 'dominance' in science and engineering is threatened — and that this a grave thing, indeed.

A cynic or a sceptic, however, is more likely to say that actually this a positive thing, because it might drive salaries up and improve conditions for PhDs. It might hurt people who rely on cheap and plentiful postdoc labour in the short term, but why should they be continually rewarded for exploitative practices anyway? And as for the perceived shortage of people willing to come to the United States — they're not to blame if a restrictive visa policy makes them consider alternative locations.

The reality is that market forces are contributing to the decline of young people taking degrees in science and industry — not just in the United States but in Europe as well. Not to worry — market forces will sort out the situation, with the best talent going to places that are willing to let them work there with a minimum amount of immigration hassle, pay them fairly and provide the infrastructure to attract and keep them there.

Paul Smaglik
Naturejobs editor



Contents

CAREERS AND RECRUITMENT

The developing role of
HIV/AIDS research p736

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

A developing theme for AIDS

Job growth in HIV research is shifting from laboratories in the developed world to the regions most affected by AIDS. Myrna Watanabe reports.

Krystn Wagner considers herself fortunate. In August 2002, having completed a medical residency and fellowship, she landed her dream job working on HIV/AIDS. She now combines medical-director duties at the Nathan Smith Clinic, a dedicated HIV clinic at Yale-New Haven Hospital, and an assistant professorship in infectious diseases at Yale University School of Medicine.

Once a growth area in both the United States and Europe, the job market for HIV/AIDS is now flattening out. That said, researchers with backgrounds similar to Wagner's — who has an MD from Yale and did a fellowship in infectious diseases at Johns Hopkins — will still be able to find jobs. These days, however, the jobs are more likely to involve clinical-research programmes in developing nations than doing basic research in Western countries.

This shift in emphasis is being driven by funding. The amount of money earmarked for HIV/AIDS is massive. In the United States alone, the National Institutes of Health (NIH) is dedicating more than \$2 billion to the disease next year, and a five-year congressional initiative is adding \$15 billion. The World Health Organization (WHO) will distribute \$5.5 billion, the Global Fund to Fight AIDS, Tuberculosis and Malaria will provide \$4.7 billion, and more than \$500 million will come from the Bill & Melinda Gates Foundation. Most of this cash will be spent on research, screening or treatment in developing nations. Governments are still spending money on domestic AIDS research, but the amount being spent on such programmes is no longer on the rise so, with the possible exception of postdocs, job opportunities are fairly static. And drug firms, feeling the pinch from the current economic downturn, are tending not to hire scientists to work on new HIV/AIDS drugs at the moment.

AIDS specialist Kevin Dieckhaus, of the University of Connecticut Health Center in Farmington, is an example of the shift towards developing nations. Spending a month in Uganda earlier this year, he saw firsthand both the need and opportunities for fieldwork in Africa. Dieckhaus was training Ugandan medical staff to recognize and treat AIDS as part of a programme run by the Infectious Diseases Society of

Beatrice Hahn (right) notes that recruitment for AIDS/HIV is relatively flat, although Krystn Wagner (below) still managed to land her dream job in the United States.



America. Although he wasn't paid, volunteers now receive a small stipend — as well as valuable experience that could help them to find a paid position later on.

"There's a shift in the AIDS field to putting the focus on developing countries," says Thomas Quinn of the US National Institute of Allergy and Infectious Diseases. Quinn works at Johns Hopkins School of Medicine in Baltimore, Maryland, where he is involved in its international programme for placing teams in Uganda.



Fieldwork: Kevin Dieckhaus (far right, middle row) visited Uganda to help train local physicians as part of an AIDS programme.

The jobs covered by the programme include administration and running clinical trials.

Opportunities for fieldwork are not limited to Africa. Vaccine researcher Yichen Lu of the Harvard School of Public Health has just completed a training course in Shanghai on AIDS for Chinese health personnel, funded by the US Office of AIDS Research. Seven Harvard physicians and one from Canada were paid to train 45 Chinese in ways to contain the epidemic.

China is investing \$1.2 billion to develop a public-health response system, and many hundreds of millions more are being spent by the government, foreign governments and international organizations on HIV/AIDS-related projects in the country. But China is finding it difficult to fill the positions created as part of these initiatives, says Lishi Bai, an HIV specialist at the Heilongjiang Provincial CDC in Harbin, China. There are many AIDS-prevention projects in China that need epidemiologists and project managers, Bai explains.

That said, Hong Kong's recent economic hardships have affected its ability to fund staff positions and caused some lay-offs, notes O. C. Lin, chief executive of the Hong Kong AIDS Foundation. And Yi-Ming Arthur Chen at the AIDS centre in Taiwan's National Yang-Ming University is negative about job opportunities for basic scientific researchers, clinical researchers, epidemiologists, clinicians, public-health workers and project managers on the island.

EVEN KEEL

Back in the United States, the job market is stable, if static. "We are hiring as many HIV/AIDS people now as we were hiring ten years ago," says Beatrice Hahn of the University of Alabama at Birmingham, who has been studying the aetiology of HIV. The only position she recently filled in her lab was for a postdoctoral fellow. She notes that basic scientists are not seeing an influx of money and that HIV/AIDS grants suffered a temporary cutback to pay for development of an anthrax vaccine.

Judith Lieberman of Harvard's Center for Blood Research complains that when the National Institutes of Health (NIH) increased the pay scale for postdocs and graduate students, it didn't include a concomitant rise in grant amounts, even though salaries comprise more than half of all monies spent in grants. So the amount of money available for research and staff has declined. Furthermore, since the terrorist attacks of

11 September 2001, the inability of many foreign-born scientists to obtain US government-funded fellowships is making it difficult to train people who may return to their home countries.

This doesn't mean that no one is hiring. Hahn recently hired a postdoc, and Stephen O'Brien, chief of the Laboratory of Genomic Diversity, National Cancer Institute, Frederick, Maryland, has been hiring postdocs and technicians, but no tenure-track people. Simon Wain-Hobson of the Pasteur Institute in Paris says that the French AIDS research agency funds postdocs and students and that the European Union will fund postdocs in collaborative work in AIDS.

A MIXED BAG

Where does industry stand in all of this? Brinda Emu-Parikh, who is doing a fellowship in infectious diseases at the University of California, San Francisco, notes that some people have left academia for the biotech industry. But all is not well in the industry. VaxGen, a biotech company in Brisbane, California, whose HIV vaccine just proved ineffective in phase III clinical trials in Thailand, has now laid people off, according to a spokesperson. Merck has eliminated about 4,400 positions, and Boehringer Ingelheim has just moved its HIV-research efforts to Canada and a spokeswoman says that it is concentrating its hiring in the cardiovascular field. Pfizer, on the other hand, is hiring for its HIV-product research groups in La Jolla, California, and Sandwich, UK, according to Martin Mackay, its senior vice-president for research and technology.

Even in industry, job openings are often linked to the developing world. For example, Indevus Pharmaceuticals in Lexington, Massachusetts, is creating vacancies in clinical trials of its vaginal microbicide in developing countries, says Al Profy, vice-president of preclinical development.

More money will soon be available for Western scientists to train physicians in the developing world. The WHO this month announced its '3 by 5 initiative', which aims to treat three million people in developing nations with anti-retroviral therapy by the end of 2005. The programme is earmarking about US\$290 million to fund staffing and activities in the countries — another step in shifting jobs to the areas that need it most. ■

Myrna Watanbe is a freelance science writer in Patterson, New York.

Web links

WHO 3 by 5 Initiative

► www.who.int/3by5

UNAIDS

► www.unaids.org

Global Fund to Fight AIDS, Tuberculosis and Malaria

► www.theglobalfund.org

Infectious Diseases Society of America

► www.idsociety.org

Bill & Melinda Gates Foundation

► www.gatesfoundation.org